Universidad Nacional Autónoma de México
Introducción a la ciencia de datos
Análisis del Trastorno del sueño
Fuentes Hernandez Brian Eduardo • *Abril 2025*

Introducción

Los **trastornos del sueño** comprenden un grupo de afecciones que alteran el patrón de sueño normal de una persona, afectando su descanso, salud física y bienestar emocional. Estas alteraciones pueden tener causas fisiológicas, neurológicas, ambientales o psicológicas, y su persistencia a lo largo del tiempo puede aumentar el riesgo de enfermedades cardiovasculares, metabólicas, psiquiátricas y accidentes laborales o viales [4].

Entre los trastornos más comunes se encuentran:

- **Insomnio**: dificultad persistente para conciliar o mantener el sueño.
- **Apnea obstructiva del sueño (AOS)**: interrupciones repetidas de la respiración por obstrucción de las vías aéreas [1].
- **Narcolepsia**: somnolencia excesiva diurna con episodios de sueño repentinos.
- **Síndrome de piernas inquietas**: necesidad incontrolable de mover las extremidades, especialmente en reposo.
- **Trastornos del ritmo circadiano**: desincronización entre el reloj biológico interno y el horario externo.

El diagnóstico suele requerir pruebas clínicas como la *polisomnografía*, que evalúa múltiples variables fisiológicas durante el sueño, así como cuestionarios estandarizados como el *Índice de Calidad de Sueño de Pittsburgh (PSQI)* o la *Escala de Somnolencia de Epworth* [3].

Históricamente, el tratamiento de estos trastornos ha incluido cambios conductuales, intervenciones médicas y el uso de dispositivos. En años recientes, la inteligencia artificial ha comenzado a emplearse como herramienta complementaria para la detección temprana, el diagnóstico automatizado y la personalización de terapias [2].

Propuesta de Casos de Uso con Machine Learning.

El análisis de datos mediante técnicas de *Machine Learning* ofrece múltiples posibilidades para abordar el estudio de los trastornos del sueño. Estas herramientas permiten identificar patrones, predecir riesgos y clasificar pacientes de manera automática con base en información recopilada.

Una primera aplicación consiste en construir modelos de clasificación que, a partir de datos como la edad, el sexo, el índice de masa corporal (IMC), la presión arterial y

la frecuencia cardíaca, predigan si una persona presenta o no un trastorno del sueño. Algoritmos como la regresión logística, los bosques aleatorios (*Random Forest*) o *XGBoost* pueden ser útiles para esta tarea. Es esperable que, con una buena selección de variables, el modelo logre predecir con alta precisión quiénes podrían estar en riesgo.

Otra posibilidad es utilizar métodos de agrupamiento no supervisado, como *K-Means*, para identificar grupos de pacientes con características similares. Esto puede ser útil para segmentar la población en perfiles de riesgo o para detectar patrones comunes entre quienes sufren ciertos trastornos. Por ejemplo, se podría encontrar un grupo compuesto mayoritariamente por personas con IMC elevado y presión alta, asociado a mayor prevalencia de apnea del sueño [1].

Asimismo, pueden aplicarse modelos de regresión para predecir la severidad del trastorno del sueño en una escala continua. Esto permitiría estimar qué tan afectada podría estar una persona según sus características clínicas. De esta forma, los profesionales de la salud podrían priorizar la atención de los casos más graves.

Finalmente, se puede analizar cuáles variables tienen mayor influencia en los modelos predictivos. Técnicas como la importancia de variables o los valores SHAP ayudan a interpretar los modelos y entender mejor qué factores están más relacionados con los trastornos del sueño. Esto resulta de gran utilidad para orientar campañas preventivas o mejorar el enfoque clínico [2].

Replicación del Código Base

Para comenzar el análisis, se llevó a cabo la replicación del código base proporcionado por la plataforma Kaggle. Este código sirvió como referencia para comprender la estructura y el enfoque inicial del problema. En particular, se trabajó con el conjunto de datos *Sleep health and lifestyle dataset.csv*, que incluye variables relacionadas con el sueño, actividad física, salud cardiovascular, IMC, nivel de estrés, ocupación, entre otras.

Principales pasos replicados

- **Importación de bibliotecas**: Se utilizaron herramientas de preprocesamiento como `StandardScaler`, `RobustScaler`, `OneHotEncoder`, y modelos como `LogisticRegression` y `XGBClassifier`.
- **Carga y exploración de datos**: Se revisó la estructura de las variables, tipos de datos y valores faltantes.
- **Visualizaciones iniciales**: Se generaron histogramas, violin plots y gráficos de conteo para explorar distribuciones y relaciones.
- **Transformación de datos**: Separación de variables numéricas y categóricas, escalamiento, codificación y partición en entrenamiento y prueba.
- **Balanceo con SMOTE**: Para corregir el desbalance de clases en la variable objetivo *Sleep Disorder*.

Este paso fue esencial para familiarizarse con los datos y establecer una línea base sobre la cual desarrollar mejoras.

Mejora o Adición al Análisis

Con base en la estructura del código replicado, se diseñaron diversas mejoras para enriquecer el análisis desde una perspectiva predictiva y explicativa.

1. Análisis de Importancia de Variables

Se utilizó el atributo `.feature_importances_` del modelo XGBoost para identificar las variables más relevantes en la predicción. Se visualizó un gráfico de barras que destacó a `Stress Level`, `Sleep Duration`, `Quality of Sleep` y `BMI Category` como factores clave.

2. Interpretabilidad con SHAP

Con el objetivo de hacer el modelo más transparente, se incorporaron los valores SHAP (SHapley Additive exPlanations). Estos permitieron:

- Visualizar el impacto global de cada variable.
- Explicar predicciones individuales.
- Generar confianza en entornos clínicos o sociales.

Se utilizaron gráficos como `summary_plot` y `force_plot` para ilustrar estas explicaciones.

3. Validación Cruzada Comparativa

Se implementó una validación cruzada estratificada de 5 pliegues utilizando la métrica `F1` ponderada para comparar el desempeño de todos los modelos entrenados. Esto permitió medir la robustez y consistencia de los clasificadores.

4. Análisis de Outliers

Se aplicó el modelo `IsolationForest` para detectar valores atípicos en el conjunto de datos. Esta técnica ayudó a identificar observaciones que podrían distorsionar el entrenamiento o influir en el rendimiento de los modelos.

5. Visualización con PCA

Se aplicó Análisis de Componentes Principales (PCA) para proyectar los datos a dos dimensiones y explorar visualmente la separación entre clases. Este análisis reveló una distribución adecuada tras el preprocesamiento, así como posibles agrupamientos naturales.

Comparación de Modelos

Para cerrar el análisis, se realizó una comparación formal entre los distintos modelos de clasificación implementados para predecir trastornos del sueño. El objetivo fue evaluar su rendimiento de forma objetiva y determinar si alguno presentaba una mejora significativa sobre los demás.

Modelos evaluados

- **Regresión Logística**
- **XGBoost (Extreme Gradient Boosting)**
- **Gradient Boosting Machine (GBM)**
- **K-Nearest Neighbors (KNN)**

Todos los modelos fueron entrenados y evaluados sobre el mismo conjunto de datos.

Evaluación con métricas clásicas

Las métricas empleadas fueron:

- **Accuracy**: proporción de predicciones correctas.
- **F1-score ponderado**: promedio ponderado entre precisión y recall.
- **Precisión y Recall ponderados**.
- **Matriz de confusión**.

Además, se utilizó **validación cruzada estratificada (5 folds)** para obtener una evaluación más robusta del rendimiento de cada modelo.

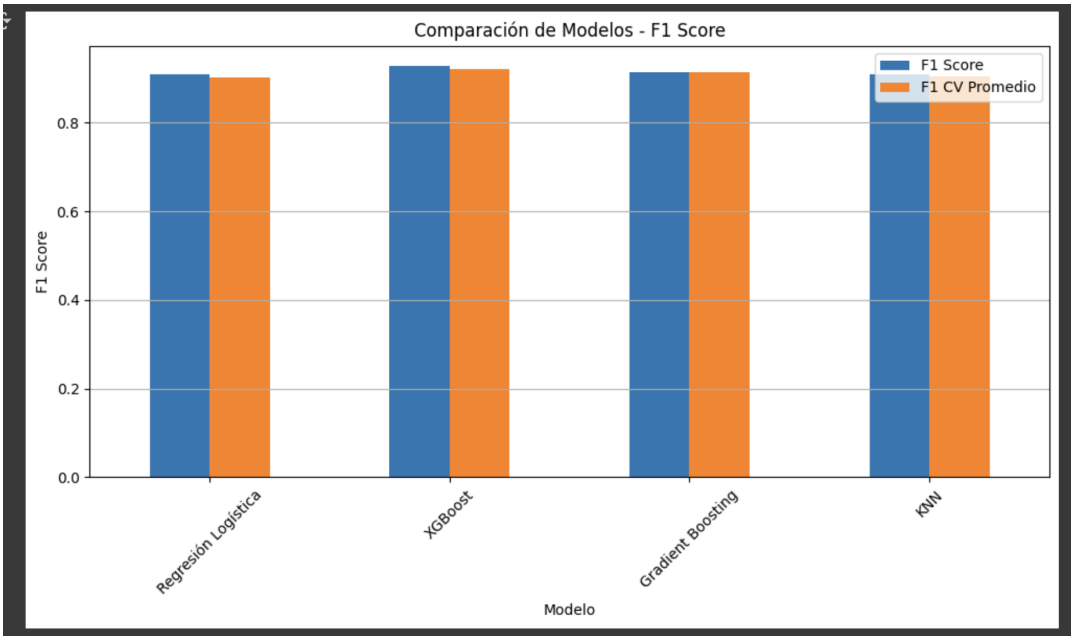

Figura 1: Comparación de F1-score promedio entre modelos

Resultados de validación cruzada

La siguiente tabla muestra el F1-score promedio y su desviación estándar calculados mediante validación cruzada:

Modelo	F1-CV Promedio	Desviación Std
Regresión Logística	0.902	0.009
XGBoost	0.922	0.015
Gradient Boosting	0.913	0.014
KNN	0.904	0.004

Cuadro 1: Resultados de validación cruzada (5 folds) por modelo

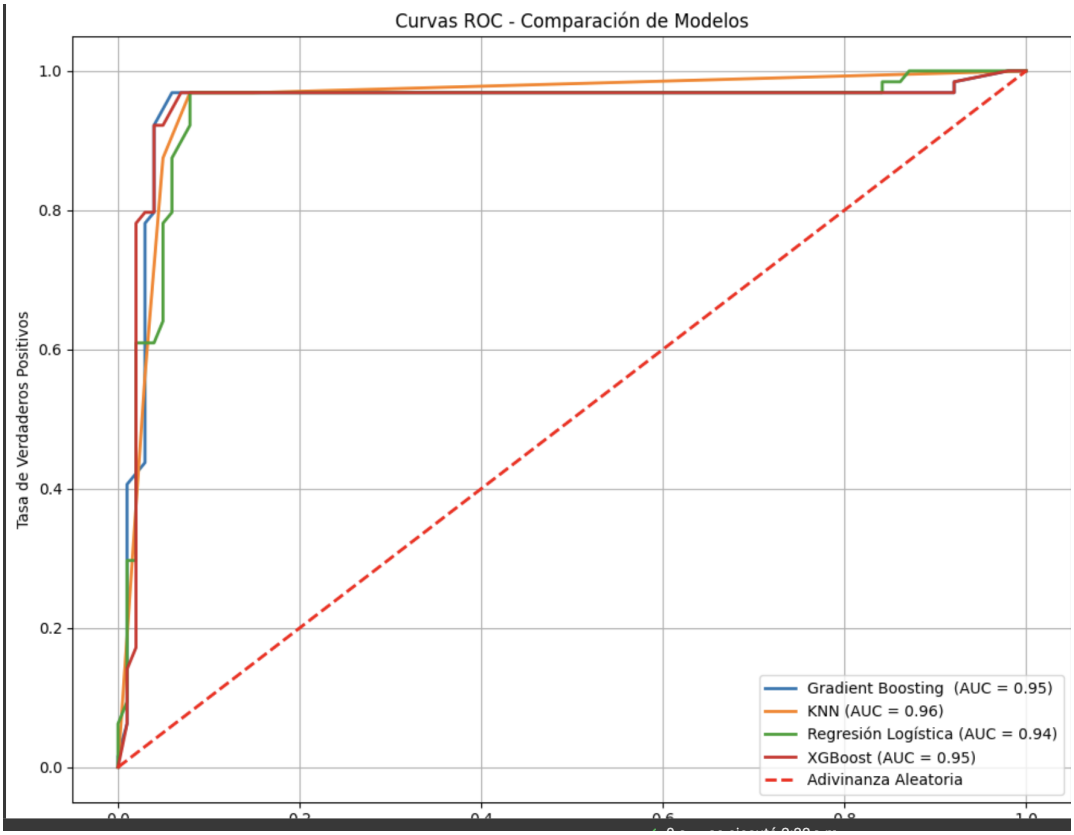

Figura 2: Curvas ROC y AUC

Se graficaron las curvas ROC para cada modelo, lo cual permitió observar su rendimiento en términos de sensibilidad vs. tasa de falsos positivos. También se calculó el área bajo la curva (AUC), que resume el comportamiento general del clasificador. Los modelos mostraron curvas similares y valores de AUC muy cercanos, lo que refuerza la idea de que su rendimiento es comparable.

A pesar de utilizar modelos complejos como XGBoost y Gradient Boosting, no se observó una mejora significativa.

Conclusión final

Los cuatro modelos analizados mostraron un rendimiento altamente competitivo y similar. Si se prioriza la simplicidad, interpretabilidad y velocidad de entrenamiento, la Regresión Logística es una excelente opción. Por otro lado, si se desea un análisis más detallado de la importancia de las variables, XGBoost combinado con SHAP proporciona explicaciones más visuales y útiles a nivel individual.

Referencias

[1] Centers for Disease Control and Prevention. Sleep and sleep disorders, 2022. Consultado el 14 de abril de 2025.

[2] Healthline. How ai is changing sleep medicine, 2024. Consultado el 14 de abril de 2025.

[3] Mayo Clinic. Sleep disorders, 2023. Consultado el 14 de abril de 2025.

[4] National Heart, Lung, and Blood Institute. What are sleep disorders?, 2023. Consultado el 14 de abril de 2025.