Dulces Sueños Predictivos

Detectando Trastornos del sueño con Ciencia de Datos

Memoria del Proyecto de Machine Learning

Naiara Saratxaga Goffard

Índice

Esta memoria presenta un análisis exhaustivo sobre la detección y el abordaje de los trastornos del sueño a través de técnicas avanzadas de ciencia de datos.

La investigación abarca la exploración y el preprocesamiento de datos, el modelado y la predicción, con el objetivo de ofrecer una solución innovadora para la identificación eficaz de estos trastornos. Asimismo, se busca comprender mejor los patrones del sueño y las anomalías que impactan en la salud y el bienestar, demostrando el potencial de la ciencia de datos para proponer futuras optimizaciones y mejorar la calidad de vida.

I Introducción

Contexto del problema y justificación de la necesidad

Los trastornos del sueño constituyen un problema significativo en la salud pública por su impacto negativo en la calidad de vida, productividad laboral y salud mental. La detección temprana y precisa es esencial para evitar complicaciones médicas y mejorar el bienestar general.
En este proyecto, propongo un sistema inteligente basado en técnicas de Machine Learning para predecir la presencia y el tipo específico de trastorno del sueño a partir de datos personales, clínicos y de estilo de vida. Con este sistema, podría facilitar intervenciones tempranas que mejoren tanto la calidad del descanso como de salud en general.

Objetivos y alcance del proyecto

Mis objetivos principales son:

1. Detectar la presencia de trastornos del sueño a partir de datos demográficos, clínicos y de estilo de vida.
2. Clasificar el trastorno detectado (insomnio, apnea del sueño, narcolepsia, etc.).
3. Identificar los factores clave relacionados con la aparición de estos trastornos.
4. Proporcionar recomendaciones prácticas y personalizadas para mejorar mi calidad de sueño.

El alcance de mi proyecto incluye:

- Desarrollar un modelo predictivo jerárquico para la detección y clasificación de trastornos del sueño.
- Segmentar a los pacientes mediante técnicas de clustering, con el fin de adaptar las recomendaciones a cada grupo.
- Crear una aplicación práctica que ofrezca recomendaciones individualizadas basadas en los resultados del modelo.

II. Dataset

Descripción del dataset que he utilizado

He empleado dos conjuntos de datos extraídos de Kaggle:

- **Sleep Health and Lifestyle Dataset,** He trabajado con un conjunto que recopila información sobre la salud del sueño y el estilo de vida de 400 personas, organizado en 13 columnas. Incluye variables como la duración y calidad del sueño, niveles de actividad física y de estrés, índice de masa corporal (IMC), presión arterial, frecuencia cardíaca, pasos diarios y la presencia de trastornos del sueño (insomnio, apnea, etc.).

- **Sleep Disorder Diagnostic Dataset:** Para el diagnóstico de trastornos del sueño, he utilizado un dataset con datos simulados que, además de la información demográfica básica (ID, edad, sexo), incluye diagnósticos de cinco tipos de trastornos del sueño y una columna binaria que indica si el diagnóstico está confirmado.

Después, fusioné ambos datasets en un único archivo (**combined_sleep_dataset.csv**) usando como variables clave "Edad" y "Género", obteniendo un total de 2.256 registros. El dataset final agrupa información demográfica, de estilo de vida y diagnóstico clínico consolidado.

Variables del dataset final:

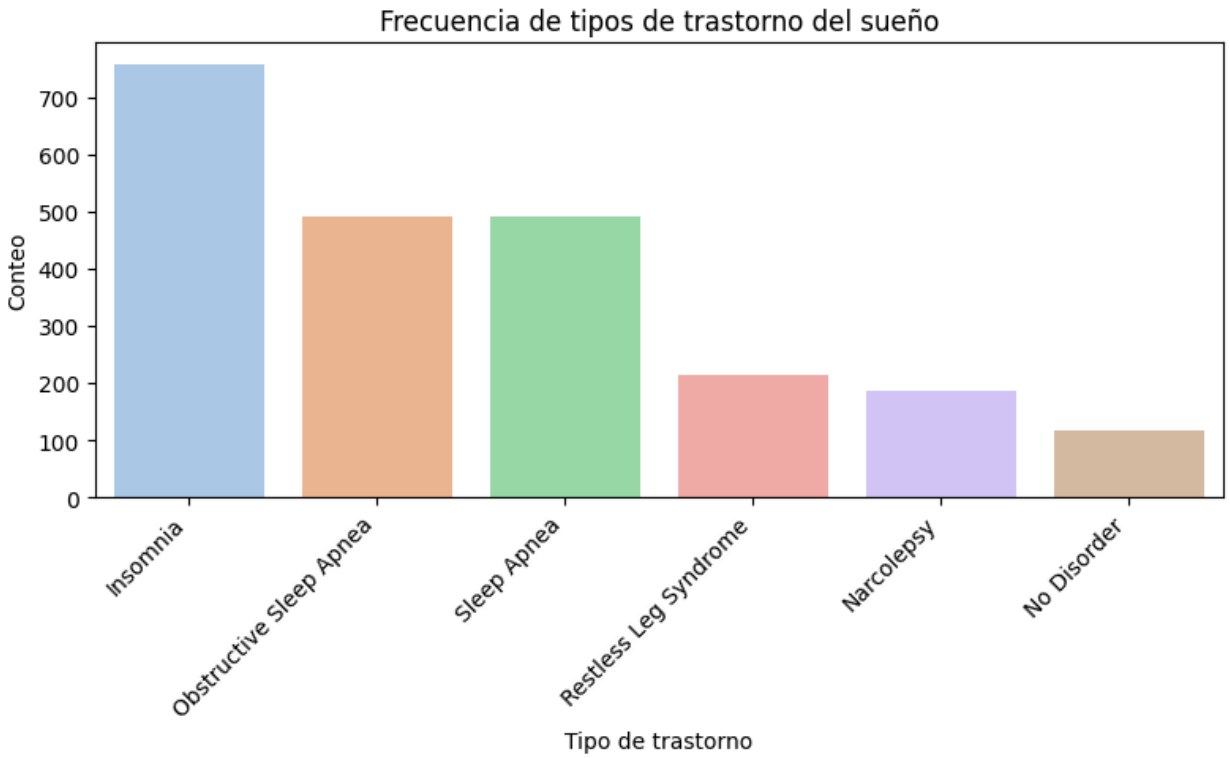
- **Género** (binario)
- **Edad**
- **Ocupación** (categórica)
- **Duración del sueño** (horas)
- **Calidad del sueño** (escala 1-9)
- **Nivel de actividad física**
- **Nivel de estrés**
- **Categoría IMC**
- **Presión arterial**
- **Frecuencia cardíaca**
- **Pasos diarios**
- **Diagnóstico confirmado** (binario)
- **Tipo de trastorno del sueño** (Insomnia, Apnea, Narcolepsia, etc.)

Análisis exploratorio de los datos (EDA)

- **Distribución de trastornos del sueño**:
    - Insomnio: 757
    - Apnea Obstructiva del Sueño: 492
    - Apnea del Sueño: 490
    - Síndrome de Piernas Inquietas: 213
    - Narcolepsia: 187
    - Sin Trastorno: 117
- **Principales correlaciones identificadas**:
    - He observado una alta correlación entre nivel de actividad física y pasos diarios (0,77).
    - Existe una correlación negativa fuerte entre estrés y calidad del sueño (–0,92).
    - La edad se correlaciona positivamente con presión arterial (0,50) y calidad del sueño (0,47).
- **Hallazgos del análisis bivariante**:
    - Los pacientes con apnea del sueño tienden a ser de mayor edad, presentar presión arterial y frecuencia cardíaca más altas.
    - Los pacientes con insomnio muestran la peor calidad y la menor duración de sueño.
- **Conclusiones clave del análisis exploratorio**:

- Variables como nivel de actividad física, edad, presión arterial y calidad del sueño resultan fundamentales para predecir la presencia y el tipo de trastorno del sueño.
- Se evidencian diferencias claras en los patrones de estilo de vida y salud entre los distintos trastornos, lo que refuerza la necesidad de un modelo predictivo personalizado.

Este análisis exploratorio me ha proporcionado una sólida base para el desarrollo del modelo predictivo y la propuesta de recomendaciones individualizadas.



- Este gráfico me hizo reflexionar sobre la importancia de aplicar técnicas adecuadas para manejar el desbalance de clases. Por ello, consideré utilizar estrategias como la ponderación de clases en los modelos supervisados, técnicas como SMOTE, y métricas de evaluación más representativas como el F1-score macro y el recall por clase, en lugar de depender únicamente de la precisión global (accuracy).

III. Preprocesamiento de los datos

Verificación de la calidad de los datos

- **Valores nulos**: Revisé la presencia de valores nulos tras fusionar los datasets y confirmé que no quedaban datos faltantes en ninguna de las variables.

Decisiones, imputaciones y transformación de variables

- Apliqué **Label Encoding** para convertir en numéricas las variables categóricas (Género, Ocupación, Categoría IMC, Presión arterial y Tipo de trastorno del sueño), de modo que pudieran integrarse sin problemas en los modelos.
- Eliminé las columnas irrelevantes o que podían inducir overfitting (por ejemplo, Patient ID, OCR_Extracted_Text, AHI_Score, SaO2_Level).
- Definí la columna de diagnóstico del trastorno del sueño como mi **variable objetivo** para el desarrollo del modelo predictivo.

Con estas transformaciones, me aseguro de que los datos estén correctamente preparados para aplicar técnicas avanzadas de Machine Learning.

IV. Modelado
Predicción de diagnóstico confirmado

En este notebook trabajé con un problema de clasificación binaria supervisada, cuyo objetivo era predecir si un paciente tenía un diagnóstico confirmado de trastorno del sueño (**Diagnosis_Confirmed**).

Para ello, entrené varios modelos supervisados:

Random Forest, en tres versiones: modelo base, con class_weight='balanced' y aplicando **SMOTE**.

XGBoost, ajustando scale_pos_weight para compensar el desbalance.

LightGBM, configurado con is_unbalanced=True.

Una Red Neuronal con **Keras**, que entrené tras aplicar **SMOT**E y ajustar los pesos de clase (class_weight).

- **Evaluación de modelos e iteraciones**

Evalué los modelos usando métricas como **precisión, recall, F1-score**, exactitud y **ROC AUC**. Este es el resumen de los resultados más relevantes:

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

En particular, con la red neuronal trabajé ajustando el **umbral de decisión** para mejorar el balance entre clases. Generé una curva de métricas **(precision, recall, F1)** para cada umbral, y encontré que el mejor resultado se obtenía en **0.35**.

Selección e interpretación del modelo final
Finalmente, seleccioné la **red neuronal con umbral 0.35** como el mejor modelo. Fue el que logró el mejor equilibrio entre clases, con un F1 macro de **0.72** y una **ROC AUC de 0.7852**, mostrando una buena capacidad discriminativa.

Este modelo logró un **recall del 73% para la clase negativa**, lo cual era una prioridad debido al fuerte desbalance del dataset. Considero que este enfoque con SMOTE, pesos de clase y ajuste de umbral fue clave para mejorar la detección de casos no diagnosticados.
Como paso adicional, también guardé un **modelo Random Forest en un pipeline con escalado** usando pickle, para integrarlo fácilmente en la aplicación final si fuera necesario.

IV. Modelado
Entrenamiento de modelos supervisados

En este segundo notebook, ***04_Entrenamiento_clasificacion_tipo_trastorno.ipynb*** me enfoqué en predecir el tipo específico de trastorno del sueño (variable Sleep_disorder), lo que planteaba un problema de clasificación multiclase supervisada.

Entrené los siguientes modelos:

Random Forest, con class_weight='balanced', **XGBoost y LightGBM**, también con ajuste de clases desbalanceadas. Un modelo **StackingClassifier**, combinando los anteriores con una regresión logística como meta-modelo. Por último, entrené también una red neuronal multicapa con **Keras**, ajustando pesos de clase y usando **one-hot encoding.**

- **Evaluación de modelos e iteraciones**

Para comparar los modelos utilicé métricas ponderadas (weighted) que penalizan correctamente los errores en clases menos representadas: **precisión, recall, F1-score y exactitud**. Estos fueron los resultados:

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

El modelo que mejor rendimiento mostró en conjunto fue **Random Forest**, con un F1 ponderado de 0.563. Sin embargo, observé que la red neuronal se acercaba mucho, sin superar claramente a los modelos ensemble.

A pesar de varios intentos con la red neuronal (ajustes en arquitectura, pesos de clase y validación temprana), el rendimiento se mantuvo similar al resto, por lo que decidí no usarla como modelo final.

Selección e interpretación del modelo final

Escogí **Random Forest** como modelo final para esta tarea de clasificación multiclase. Fue el más equilibrado, especialmente en clases mayoritarias como el insomnio (clase 0) y los pacientes sin trastorno (clase 5), donde obtuvo:

- **Precisión (clase 0)**: 0.90

- **Recall (clase 5)**: 0.93

Sin embargo, las clases minoritarias (como narcolepsia o síndrome de piernas inquietas) mostraron peores resultados en recall y F1, lo que confirma el reto del desbalance. La matriz de confusión también reflejó este comportamiento, con más confusión entre clases intermedias.
Finalmente, guardé el modelo Random Forest con su preprocesamiento completo en un **pipeline serializado con pickle**, listo para su uso en producción.

IV. Modelado
Entrenamiento de modelos no supervisados En este tercer bloque, ***05_Entrenamiento_segmentacion_pacientes.ipynb***, enfoqué mi trabajo en **agrupar pacientes sin usar una variable target**, utilizando técnicas de **clustering no supervisado** para descubrir patrones naturales en los datos.

Probé tres métodos principales:

- **K-Means** (con selección del número óptimo de clusters mediante el método del codo y el índice de silueta)
- **DBSCAN** (útil para detectar ruido y formas arbitrarias)
- **Clustering jerárquico aglomerativo** con el método de Ward y dendrogramasAntes del modelado, apliqué un escalado **StandardScaler** a las variables numéricas y reducción con **PCA** a 2 componentes para facilitar la visualización.

- **Evaluación de modelos e iteraciones**
Para **K-Means**, exploré valores de k desde 2 hasta 10. El **método del codo** y la **silueta promedio** sugirieron que **k = 4** ofrecía un buen equilibrio entre compactación y separación.DBSCAN, por el contrario, generó **33 grupos**, muchos de ellos interpretados como ruido (-1), lo que me llevó a descartarlo para el análisis principal. El clustering jerárquico con k = 4 permitió una segmentación coherente y se usó como apoyo visual y comparativo.

Selección e interpretación del modelo final

Elegí **K-Means con k=4** como modelo final por su interpretabilidad, estabilidad y coherencia con el análisis PCA. Apliqué este modelo al dataset y generé un perfil detallado para cada grupo:

Cluster	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

Cada grupo muestra un **perfil claramente diferenciable**, lo que valida el uso de clustering como herramienta complementaria al modelado supervisado.
Finalmente, guardé el pipeline de segmentación completo (escalado + PCA + KMeans) con **pickle** en **sleep_patient_segmentation.pkl**, dejándolo listo para reutilización en la aplicación práctica.

V. Predicción y Resultados Finales

1. **Diagnóstico binario** (trastorno del sueño sí/no) con una red neuronal optimizada por umbral (F1=0.73).
2. **Clasificación del tipo de trastorno** usando un Random Forest balanceado (F1 ponderado ≈ 0.56), que identifica bien los casos claros como “Sin trastorno” o “Apnea”, aunque tiene dificultades con clases menores como “Narcolepsia”.
3. **Segmentación no supervisada** con K-Means (k=4), que revela perfiles diferenciados de pacientes según edad, calidad del sueño, actividad física y salud general.

Visualización de los resultados finales y predicciones

- Se usaron **matrices de confusión** y **reportes de clasificación** para evaluar cada modelo.
- El modelo de diagnóstico binario mejoró su rendimiento con **SMOTE y ajuste del umbral**, consiguiendo una mayor sensibilidad en la clase minoritaria.
- Los resultados del clustering se visualizaron mediante **PCA**, mostrando una separación clara entre perfiles, lo que respalda su uso para adaptar recomendaciones.

- Todos los modelos se empaquetaron como **pipelines y se guardaron con Pickle**, listos para desplegarse en una aplicación interactiva.

He creado una aplicación interactiva con Streamlit que permite introducir datos personales, predecir trastornos del sueño, clasificarlos y mostrar el perfil del usuario de forma visual y comprensible.

VI. Conclusiones y futuros pasos

Tras implementar y evaluar varios modelos, he comprobado que es factible detectar y clasificar trastornos con un rendimiento razonable empleando datos clínicos y de estilo de vida.La combinación de algoritmos supervisados ***(Random Forest, XGBoost y redes neuronales)*** junto con técnicas de balanceo de clases (**SMOTE** y ajuste de **class_weight**) ha reforzado especialmente la detección de casos negativos, mitigando el sesgo derivado del desbalance en las etiquetas.

Entre los **puntos fuertes** de este proyecto destaco:

- **Integración de enfoques heterogéneos**, abarcando clasificación binaria, multicategórica y clustering.
- **Análisis exhaustivo de métricas clave**, que permite comprender en detalle el comportamiento de cada modelo.
- **Desarrollo de una aplicación interactiva**, que facilita la interpretación de resultados por parte del usuario final.

Sin embargo, identifico también **áreas de mejora**:

- La **naturaleza sintética** y la **origen limitado** de los datos actuales, que pueden no reflejar por completo la variabilidad clínica real.
- Un **rendimiento subóptimo** en las clases minoritarias, que exige ajustes más finos de hiperparámetros o estrategias de muestreo.

Pasos futuros recomendados

1. **Ampliar y diversificar** el dataset con fuentes reales (bases de datos de hospitales o clínicas).
2. **Incorporar nuevas variables** (historial médico, hábitos nocturnos, etc.) que añadan riqueza informativa al modelo.
3. **Explorar arquitecturas avanzadas**, como redes neuronales recurrentes o transformers, para capturar dependencias temporales.
4. **Mejorar la interfaz de la aplicación**, dotándola de generación automática de informes personalizados e integración con dispositivos de monitoreo del sueño.

muchas gracias