Análisis del Transtorno del Sueño

Zahid Medrano Flores
zahidmedrano@ciencias.unam.mx

Temas Selectos en Biomatemáticas. **Introducción a la Ciencia de Datos.**.
Facultad de Ciencias. UNAM. Semestre 2025-2.

Resumen

Este reporte aborda el problema de los trastornos del sueño a través del análisis de un dataset específico, buscando no solo explorar las características y patrones presentes en los datos, sino también proponer y evaluar la aplicación de técnicas analíticas y modelos de Machine Learning como herramientas potenciales para arrojar luz sobre este importante desafío de salud.

1. Introducción

El sueño es una parte muy importante de la vida y salud humana, es tan vital y necesario como una buena nutrición o actividad física regular. En este breve periodo que tenemos nuestro cuerpo y mente se recuperan, se consolida la memoria, se regula el estado de ánimo y algunas otras funciones esenciales que tenemos para el bienestar integral. Sin embargo, a nivel mundial se presentan transtornos del sueño, los cuales son condiciones que afectan la calidad del mismo.

Los trastornos del sueño no son simplemente una molestia; representan un problema de salud pública creciente con muchas implicaciones a nivel individual y social. La falta crónica de sueño o un sueño de mala calidad está asociado con un mayor riesgo de desarrollar una amplia gama de problemas de salud, incluyendo enfermedades cardiovasculares, diabetes, obesidad, trastornos de salud mental (como depresión y ansiedad) y una disminución inmunológica. Además, el impacto llega a nuestra vida diaria, afectando nuestra concentración, toma de decisiones o rendimiento en general.

La complejidad de los trastornos del sueño se debe a su diversa naturaleza y en la variedad de agentes que pueden contribuir a ellos, desde hábitos de vida y condiciones médicas subyacentes hasta factores genéticos y ambientales. Tradicionalmente, el diagnóstico y manejo se han basado en la evaluación clínica, cuestionarios y estudios especializados. Sin embargo, la creciente disponibilidad de datos sobre patrones de sueño, estilos de vida y salud, así como los avances en ciencia de datos y Machine Learning, abren nuevas puertas para comprender mejor estos trastornos, identificar poblaciones en riesgo y desarrollar mejores enfoques para su detección y tratamiento.

2. Desarrollo

En un proyecto de Ciencia de Datos, el objetivo principal es construir modelos matemáticos y computacionales que tienen la capacidad de "aprender" a partir de datos existentes, para poder realizar predicccciones sobre datos nuevos.

2.1. Análisis exploratorio

En un análisis exploratorio se busca analizar y de alguna forma, manipular nuestro conjunto de datos, para poder obtener mejores resultados en nuestros modelos predictivos.

De manera específica, primero se busca la "forma" general que tienen los datos, así como, las caracteristicas que presentan; realizamos una busqueda rápida de como se ven los datos, su desviación, distribución, etc.

El proceso fue el siguiente:

1. Carga de datos. Se carga el dataset y se realiza un análisis exploratorio inicial, donde entendemos

la estructura de los datos, las variables y la distribución de clases (los transtornos del sueño) presentes.

2. Limpieza. Se tratan los valores nulos, se corrigen errores y las varias inconsistencias que se presenten a lo largo del dataset.

3. Transformación. En esta parte, realizamos distintos procesos para transformar los datos. Se realiza un manejo de variables categóricas, una normalización o escalado de las variables numéricas, entre otros procesos.

4. Oversampling. Si nuestro dataset es muy pequeño, y lo vemos necesario, se realiza un sobremuestreo para aumentar el número de ejemplos de la clase más pequeña en un conjunto de datos desequilibrado. Esto es opcional, pero muchas veces ayuda.

2.2. Split

A partir de nuestro dataset completo, se realiza una división de nuestro conjunto para el entrenamiento (train) y las pruebas o inferencia (test). Una división común es de 70 / 30 para el entrenamiento y el test respectivamente.

2.3. Modelos

Hay una gran variedad de modelos, los cuales se escogen de acuerdo al problema que se esté atacando. En nuestro caso se trata de un proyecto enfocado en la clasificación de transtornos del sueño. Por lo tanto tenemos algunas opciones.

- **Logistic Regression:** Es un algoritmo de clasificación lineal que modela la probabilidad de que una instancia pertenezca a una clase particular. Funciona aplicando una función logística (sigmoide) a una combinación lineal de las características de entrada. [1]
- **Gradient Boosting Machine (GBM):** Es una técnica de boosting que construye un modelo predictivo en etapas, y generaliza la optimización de la función de pérdida utilizando el descenso de gradiente. [2]
- **eXtreme Gradient Boost:** Es una implementación optimizada y muy eficiente del algoritmo Gradient Boosting. Se destaca por su velocidad, rendimiento y capacidad para manejar diversos tipos de datos y problemas. Construye una serie de árboles de decisión de forma secuencial, donde cada nuevo árbol intenta corregir los errores cometidos por los árboles anteriores, pero con optimizaciones que lo hacen muy robusto y rápido.
- **K-Nearest Neighbors (KNN):** Para clasificar un nuevo punto de datos, KNN encuentra los 'k' puntos de datos más cercanos en el espacio de características (basado en una métrica de distancia) y asigna la clase más común. [3]
- **C-Support Vector (SVC):** Es un tipo de SVM utilizada para tareas de clasificación. Encuentra el hiperplano óptimo que mejor separe las diferentes clases en el espacio de características, maximizando el margen entre los puntos de datos de las clases más cercanas a ese hiperplano (los vectores de soporte). [4]
- **Random Forest:** Es un algoritmo de aprendizaje por conjunto (ensemble learning) que funciona construyendo un gran número de árboles de decisión durante la fase de entrenamiento. Para la clasificación, combina las predicciones de todos los árboles para determinar la clase. [5]

3. Resultados

En esta sección se presentan los resultados obtenidos para cada uno de los modelos que se probaron.

3.1. Logistic Regression

La complejidad de implementación de este modelo no es tan grande, no tiene tantos hiperparámetros que se puedan cambiar de manera rápida y que sean de gran impacto. Por lo tanto el modelo es bastante fácil de aplicar.

A continuación se presenta la figura 1 con la matriz de confusión para poder observar el comportamiento que presenta el modelo al momento de predecir los transtornos del sueño.

3.2. Gradient Boosting Machine

De manera similar, la implementación del este modelo se realizó de manera simple. De hiperparámetros no hay mucho que cambiar, pero se utilizaron mas n - estimadores para probar el comportamiento que presentara.

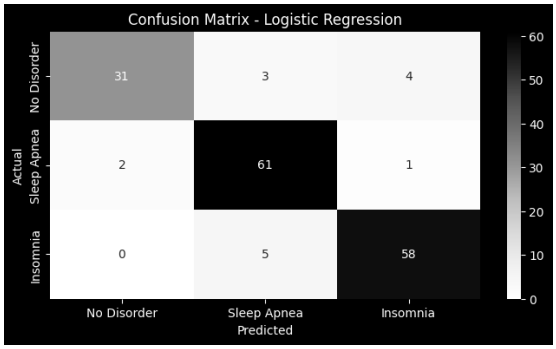

Figura 1. Matriz de confusión para el modelo de Logistic Regression.

La figura 2 que se presenta a continuación, muestra la matriz de confusión, con la cual vemos que tan bien predijo nuestros transtornos.

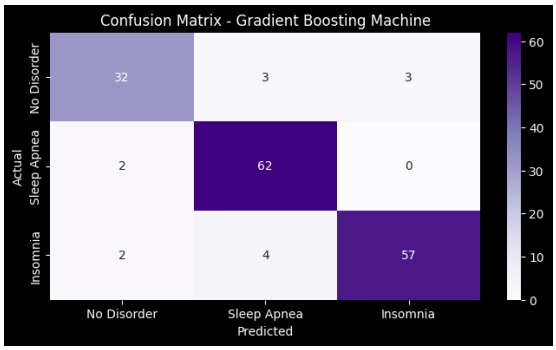

Figura 2. Matriz de confusión para el modelo GBM.

3.3. eXtreme Boosting Machine

3.4. K-Nearest Neighbors

3.5. C-Supoort Vector

3.6. Random Forest

4. Análisis y Discusión

5. Conclusiones

Referencias

[1] Cramer, J. S. (2002). The Origins of Logistic Regression. *TI.*

[2] Friedman, J. H. (1999). Greedy Function Approximation: A Gradient Boosting Machine.

[3] Cunningham, P., & Delany, S. J. (2021). k-Nearest Neighbour Classifiers - A Tutorial. *ACM Computing Surveys.*

[4] Novakovic, J., & Veljovic, A. (2011). C-Support Vector Classification: Selection of Kernel and Parameters in Medical Diagnosis. *IEEE 9th International Symposium on Intelligent Systems and Informatics.* https://doi.org/doi:10.1109/sisy.2011.6034373

[5] Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random Forests. *Ensemble Machine Learning*, 157-175. https://doi.org/http://dx.doi.org/10.1007/978-1-4419-9326-7_5

Apéndice A. Test