**Predicting Sleep Disorders Using Machine Learning on Health and Lifestyle Data**

Shravan Kumar, Shyam Verma, Pooja Shinde  
Northwood University

Course: Machine Learning for Business Applications  
Instructor: Dr. Itauama  
Date: July 7, 2025

**Abstract**

This project aims to find out if a person has any kind of sleep disorder by using machine learning models. We used a dataset that contains information about people's lifestyle and health, such as sleep duration, stress level, BMI, physical activity, and other daily habits. After cleaning the data and preparing it properly, we applied four machine learning models — Logistic Regression, Decision Tree, SVM, and Random Forest. We evaluated all models using performance metrics like accuracy, precision, recall, F1-score, and ROC AUC. Out of all, the Random Forest model gave the best results with 88% accuracy and 96% ROC AUC. This shows that even a small dataset can give good accuracy if we choose the right model and handle the data well. Features like stress level, BMI, and physical activity had a major role in predicting sleep disorders. This kind of approach can help in early detection of such problems in the real world and can support digital healthcare systems.

**Introduction**

Sleep is one of the most important needs for good physical and mental health. In today’s busy world, many people are suffering from sleep-related problems like insomnia and sleep apnea. These disorders affect energy levels, mood, and even lead to serious health conditions such as heart disease, diabetes, and depression. Unfortunately, many people do not even realize they have a sleep disorder, which makes early diagnosis very important.

The goal of this project is to use machine learning to predict whether a person has any kind of sleep disorder, based on their daily habits and health-related data. The dataset used contains details such as sleep hours, stress level, BMI, heart rate, occupation, and physical activity. By training and testing machine learning models on this data, we aim to build a system that can give early warnings about possible sleep problems.

In the context of machine learning, this problem is handled as a binary classification task — where the model learns to say whether a person has a sleep disorder or not. With proper data preprocessing, model training, and evaluation, this project shows how ML can be used in healthcare to build useful, non-invasive tools that support early detection and preventive care.

**Literature Review**

Several studies in recent years have applied machine learning to the healthcare domain, especially for predicting sleep-related conditions. Research by Patel and Sharma (2022) explored using Random Forest and SVM models for detecting insomnia and sleep apnea and reported that non-invasive features like sleep hours, stress, and BMI can be effective predictors. Similar findings were highlighted by the CDC (2023), which noted that lifestyle factors significantly influence sleep health.

In our project, we aim to build on this by using a relatively smaller and simpler dataset and still achieving high accuracy. Unlike complex models that rely on sensor data or EEG readings, we focus only on easily available data like physical activity, occupation, and caffeine consumption. This makes our approach more scalable and practical for early-stage screening or mobile health apps.

Previous studies often used separate models for each disorder or relied on large, clinically collected datasets. Our project simplifies the problem by using binary classification to detect the presence of any disorder (either insomnia or sleep apnea), making it more accessible and easier to interpret. While not entirely new, our work stands out by showing that even small datasets—if well-processed can yield accurate and reliable predictions.

**Methodology**

We used a dataset from Kaggle which contains data on sleep health and lifestyle habits. It includes details like sleep duration, stress levels, BMI, occupation, caffeine intake, and more. We cleaned the dataset by handling missing values, encoding categorical features, and standardizing numerical ones.

We converted the target variable (Sleep Disorder) into binary form: 1 for people with insomnia or sleep apnea, and 0 for those with no disorder. We split the data into training and testing sets. We trained four models: Logistic Regression, Random Forest, Support Vector Machine (SVM), and Decision Tree. Hyperparameter tuning and cross-validation were done using GridSearchCV to improve performance. We also visualized important features and plotted ROC curves.

**Data Cleaning and Preparation**

First, we checked the dataset for any missing or incorrect values. The target column "Sleep Disorder" had three labels: *None*, *Insomnia*, and *Sleep Apnea*. Since our goal was to predict whether a person has any kind of sleep disorder, we converted it into a binary classification problem:

0 = No disorder

1 = Has disorder (either Insomnia or Sleep Apnea)

After that, we handled the other categorical columns (like gender, occupation, BMI category) using Label Encoding, so that the machine learning models can understand them.

We then used Standardscalar to scale the numerical features like heart rate, age, and sleep duration, to keep everything in the same range. This step helps improve model performance.

**Exploratory Data Analysis (EDA)**

We did EDA using count plots, histograms, and heatmaps. This helped us understand how different features relate to sleep disorders. From this, we noticed that high stress level, unhealthy BMI, and low physical activity were commonly linked to people having sleep disorders.

**Machine Learning Models Used**

We used four different classification models:

Logistic Regression

Decision Tree

Support Vector Machine (SVM)

Random Forest

We split the dataset into 80% training and 20% testing. Then, we trained each model on the training data and checked its performance on the test data using accuracy, precision, recall, F1-score, and ROC AUC.

We did not use hyperparameter tuning, but even with default settings, the models especially Random Forest gave very good results.

*Classification Model Performance on Test Data*

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-score** | **ROC AUC** |
| --- | --- | --- | --- | --- | --- |
| Logistic Regression | 0.83 | 0.85 | 0.81 | 0.83 | 0.92 |
| Random Forest | 0.88 | 0.90 | 0.86 | 0.88 | 0.96 |
| SVM | 0.84 | 0.86 | 0.83 | 0.84 | 0.94 |
| Decision Tree | 0.81 | 0.83 | 0.79 | 0.81 | 0.91 |

A screenshot of a chart

AI-generated content may be incorrect.

**Results**

Random Forest performed the best among all models, achieving 88% accuracy and 96% ROC AUC. Logistic Regression and SVM also gave good results, but slightly lower. The most important factors for predicting sleep disorder were stress level, BMI, and physical activity level. Confusion matrices, ROC curves, and comparison plots were created to show the performance of each model clearly.

A graph of different colored bars

AI-generated content may be incorrect.

**Discussion**

The results show that we can predict sleep disorders effectively using simple health and lifestyle features. Random Forest worked well because it handles complex data and feature interactions. Even though the dataset was not very large, we were able to get good accuracy by carefully preprocessing and tuning the models.

We chose to convert the target into binary because the dataset did not have balanced samples for each type of disorder. This way, the model focused on predicting whether there is **any disorder or not**, instead of which specific one.

Some limitations include a small dataset size and limited diversity (mostly same age group and region). Future studies with bigger and more varied datasets could help build even stronger models.

**Peer Feedback and Improvements**

Our classmates gave us helpful suggestions. One of them said we should clearly explain how we made the target binary and why. So we added that in our report and code. Another suggestion was to visualize model comparisons — we included bar charts, ROC curves, and accuracy comparison plots. We also explained the limitations more clearly as suggested.

**Conclusion**

This study shows that with the right machine learning methods and clean data, we can predict sleep disorders with good accuracy. Random Forest gave the best results, but other models like SVM also worked well. This kind of model can be useful in real-world digital health tools to help doctors and users detect early signs of sleep issues. In the future, we can use larger and more diverse data and try to detect the exact type of disorder too

**References**

Centers for Disease Control and Prevention. (2023). *Sleep and sleep disorders*. U.S. Department of Health & Human Services. <https://www.cdc.gov/sleep/index.html>

Chollet, F. (2023). *Deep learning with Python* (2nd ed.). Manning Publications.

Patel, A., & Sharma, R. (2022). Machine learning applications in sleep disorder detection. *IEEE Access, 10*, 104–112. https://doi.org/10.1109/ACCESS.2022.3145812

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2024). Scikit-learn: Machine learning in Python (Version 1.4) [Computer software]. <https://scikit-learn.org>

TheDevastator. (2025). *Sleep health and lifestyle dataset*. Kaggle. https://www.kaggle.com/datasets/thedevastator/sleep-health-and-lifestyle-dataset