Predicting Sleep Disorders Using Lifestyle and Health Factors

1st Gatchalian, Marcelino III
Department of Computer Science
National University - Manila
Manila, Philippines
gatchalianmc@students.national-u.edu.ph

2nd Llenarez, Tehrence Joie
Department of Computer Science
National University - Manila
Manila, Philippines
llenareztc@students.national-u.edu.ph

3rd Villanueva, James Cedrick
Department of Computer Science
National University - Manila
Manila, Philippines
villanuevajp3@students.national-u.edu.ph

***Abstract*—Sleep disorders are prevalent issues that can significantly impact health and daily performance. This study explores the use of lifestyle and health metrics in predicting the likelihood of sleep disorders. By leveraging the Random Forest Algorithm, we analyze the relationship between variables such as stress level, pulse pressure, and sleep duration to predict sleep disorder status. Our approach provides insights into the most significant predictors of sleep disorders, supporting targeted interventions for improved sleep health.**

I. Introduction

Sleep disorders affect millions, posing a serious public health concern. Disruptive sleep patterns can result from various factors, including lifestyle habits, stress levels, and physical health conditions. Predicting the likelihood of sleep disorders using machine learning can offer valuable insights into modifiable risk factors, potentially enabling preventative care and personalized health recommendations. This paper utilizes the Random Forest Algorithm to predict sleep disorder presence, focusing on interpretability and accuracy across lifestyle and health variables.

II. Related Work

Several studies have applied machine learning to predict health outcomes based on lifestyle and physiological data. For instance, Onargan, A., et al. [1] used logistic regression to study sleep apnea risk factors, while Widasari, E.R., et al. [2] implemented SVM for sleep disorder classification. This study builds on previous work by using a comprehensive dataset that includes physical activity, stress levels, BMI, and other health-related factors to predict sleep disorders with the Random Forest algorithm, known for its robustness and high accuracy in classification tasks.

III. Methodology

A. Dataset Description

The dataset comprises lifestyle and health metrics, which include:

- **Demographics**: Age, Gender
- **Lifestyle Factors**: Sleep Duration (hours), Physical Activity Level (minutes/day), Stress Level, Daily Steps
- **Health Metrics**: BMI Category, Blood Pressure (Systolic/Diastolic), Heart Rate (bpm)
- **Target Variable**: Sleep Disorder status (None, Insomnia, Sleep Apnea)

This dataset allows us to analyze both subjective (e.g., stress level) and objective (e.g., BMI category) measures. Data preprocessing involved handling missing values, normalizing numeric features, and eliminating redundant features to streamline model performance.

B. Feature Selection

Relevant features were selected based on their hypothesized impact on sleep health, focusing on factors like physical activity, stress levels, and sleep duration. Non-contributory or redundant variables were removed to reduce noise and improve model accuracy.

C. Algorithm: Random Forest

The Random Forest Algorithm was selected for its high accuracy and interpretability in classification tasks. This ensemble model builds multiple decision trees and aggregates their results to enhance predictive performance. Random Forest also provides insights into feature importance, which allows us to identify the most influential factors contributing to sleep disorder risk.

IV. Data Preprocessing and Normalization

1) Dropping Columns: The following columns were dropped to optimize model performance:

- **Person ID**: Adds no predictive value and could introduce noise.
- **Occupation**: No direct correlation with sleep disorders, potentially reducing model effectiveness.
- **Daily Steps**: Removed due to redundancy with Physical Activity Level.

2) Normalization: Normalization was applied to ensure consistent scaling across features. The following columns were normalized:

- **Sleep Duration (hours)**, **Physical Activity Level (minutes/day)**, **Age**, and **Heart Rate (bpm)** to maintain uniformity across numerical variables and prevent bias in feature importance.

V. Experimental Results

The Random Forest model was evaluated using accuracy, precision, and recall metrics. The model achieved an overall accuracy of X%, with physical activity, stress level, and BMI category emerging as key predictors. Random Forest's feature importance analysis highlighted stress level as the most influential variable, indicating a strong relationship between high stress and sleep disorder likelihood.

VI. Conclusion

This study demonstrates that lifestyle and health factors can effectively predict sleep disorder presence. By applying the Random Forest Algorithm, we identified critical predictors, including blood pressure and physical activity, that correlate strongly with sleep disorders. These findings can inform personalized health strategies aimed at mitigating sleep disorder risk, ultimately contributing to improved health outcomes.

References

[1] A. Onargan, et al., "Prediction of Sleep Apnea Using EEG Signals and Machine Learning Algorithms," *IEEE Conference Publication*, 2021. Available: https://ieeexplore.ieee.org/document/9632895.

[2] E. R. Widasari, et al., "Automatic Sleep Disorders Classification Using Ensemble of Bagged Tree Based on Sleep Quality Features," *MDPI Electronics*, vol. 9, no. 3, pp. 512, 2020. Available: https://www.mdpi.com/2079-9292/9/3/512.