

---

# Predicting Sleep Disorders Using Supervised Machine Learning on Lifestyle Data.

Department of Information Technology , Madras Institute of Technology , Anna University, Chennai  
e-mail: abc@gmail.com  
submitted to : DR J Dhalia Sweetlin

## ABSTRACT :

Sleep disorders affect millions worldwide, impacting health and quality of life. Early detection can help reduce health risks and medical costs. This study presents a data-driven approach to predict sleep disorders using the Sleep Health and Lifestyle Dataset, which includes 400 records with features like age, occupation, sleep duration, and physical activity. Our method involves data cleaning, analysis, feature selection, and machine learning model development. We tested several algorithms including Logistic Regression, Linear Support Vector Machines, Random Forest, and Gradient Boosting. Among these, **Logistic Regression achieved the highest accuracy**, making it the most effective model for classifying sleep disorders into None, Insomnia, and Sleep Apnea. The models were evaluated using accuracy, precision, recall, and F1-score. The results show that machine learning can accurately predict sleep disorders using lifestyle data. This approach can help healthcare providers and insurance companies in early screening and decision-making.

## INDEX TERMS :

Sleep disorder prediction, machine learning, logistic regression, Linear supervised learning, data science, healthcare analytics, classification models, lifestyle data.

## I. INTRODUCTION

Sleep is a critical part of human life, necessary for both physical recovery and mental performance. It helps the body rest, strengthens the immune system, and allows the brain to organize and store memories. Without proper sleep, people may suffer from reduced concentration, poor memory, mood swings, and serious long-term health problems like heart disease, diabetes, and obesity.

The quality of sleep plays a major role in daily functioning, especially for children and elderly individuals. For example, older adults who don't get enough quality sleep are more likely to have accidents, especially while driving. Despite the importance of sleep, many people worldwide continue to struggle with sleep disorders, which prevent them from getting the rest their bodies and minds need.

A sleep disorder is a medical condition that affects the ability to fall asleep, stay asleep, or achieve restful sleep. Common types of sleep disorders include insomnia (trouble sleeping), sleep apnea (breathing interruptions during sleep), narcolepsy (sudden sleep attacks), and other conditions such as sleepwalking. These problems can lead to daytime fatigue, stress, and reduced performance at work or school. In serious cases, untreated sleep disorders can

cause life-threatening conditions.

Traditionally, diagnosing sleep disorders involves manual evaluations by medical professionals using a process called polysomnography (PSG). This test monitors brain activity, heart rate, breathing, and other physical signs during sleep. However, reading PSG data is complex and time-consuming, and results can vary depending on the doctor's interpretation. Manual analysis is also prone to human error. As a result, many sleep disorders remain undiagnosed or are detected too late.

A global survey conducted by Philips for World Sleep Day in 2021 showed just how widespread this issue is. Over 13,000 adults across 13 countries were surveyed. The results revealed that only 55% were satisfied with their sleep. The COVID-19 pandemic had a major impact on sleep quality, with 37% of people reporting that the pandemic made it harder to sleep. Additionally, 37% of participants reported suffering from insomnia, 29% snored, 22% had shift-work sleep disorders, and 12% experienced sleep apnea. These numbers highlight the need for more efficient and accessible ways to identify sleep problems.

---

Sleep occurs in five main stages: Wakefulness, N1, N2, N3, and REM (Rapid Eye Movement). Wakefulness is when the person is fully alert. In N1, sleep begins lightly, and brain activity starts to slow. N2 and N3 represent deeper stages of sleep, with N3 being the deepest, where it's very hard to wake someone. REM is a stage where the brain is very active, and dreams usually occur. Each stage has a specific function for body repair, brain processing, and memory storage.

Because PSG studies are expensive and not available to everyone, researchers have begun using data science and machine learning to find more affordable ways to detect sleep disorders. Machine learning can analyze large amounts of data quickly and learn patterns that help predict whether a person is likely to have a sleep disorder. These models can use simple, everyday data — such as age, sleep duration, physical activity, occupation, and stress levels — to make accurate predictions.

This project aims to use machine learning to predict sleep disorders using a public dataset called the Sleep Health and Lifestyle Dataset. The dataset contains records of around 400 individuals and includes information on their age, gender, occupation, hours of sleep, quality of sleep, stress level, BMI category, and whether they have a sleep disorder. The goal is to build a model that can classify people into three categories: no disorder, insomnia, or sleep apnea.

By applying machine learning to this lifestyle data, we aim to support early identification of sleep disorders in a faster, low-cost, and non-invasive way. This can be helpful not just for healthcare providers, but also for insurance companies looking to assess risks, and even for individuals using mobile health apps to track and manage their sleep habits.

In short, this paper presents a modern and practical approach to identifying sleep disorders using artificial intelligence and simple lifestyle data. It highlights the growing potential of data-driven healthcare in predicting and preventing health issues before they become serious.

## 1. RELATED WORK:

Over the past several years, many researchers have worked on developing machine learning and deep learning systems to identify and classify sleep disorders. Traditional methods such as polysomnography (PSG), although effective, are expensive and require experts to interpret the results. As a result, researchers are exploring how machine learning algorithms (MLAs) can help automate and improve the accuracy of sleep disorder diagnosis using different types of data such as ECG and EEG signals.

One study by **Salari et al.** focused on detecting sleep apnea using ECG (electrocardiogram) signals with the help of ML algorithms like Support Vector Machine (SVM) and Random Forest (RF) [1]. These models were used to

analyze heartbeat patterns and identify signs of sleep apnea. Although they were successful in detecting apnea with decent accuracy, the researchers pointed out some key challenges. First, ECG signals can vary a lot between people, which makes it harder to build models that work well for everyone. Second, high-quality ECG datasets are not always available for training machine learning models, which limits the accuracy and generalization of the results.

Another study by **Li et al.** used EEG (electroencephalogram) spectrograms to classify sleep stages using deep learning techniques [2]. They applied a combination of convolutional layers and bidirectional Long Short-Term Memory (LSTM) layers to learn both the frequency and time features from the EEG data. This hybrid approach allowed their model to better understand sleep patterns across different stages, such as light sleep, deep sleep, and REM sleep. Their model achieved high accuracy, but they also noted that the dataset used was unbalanced, which means that some sleep stages had fewer examples than others. This imbalance affected the model's ability to perform equally well across all sleep stages.

In another study, **Han and Oh** predicted the severity level of Obstructive Sleep Apnea (OSA) using real patient data collected from over 4,000 individuals [3]. They used both supervised learning algorithms, such as gradient boosting and random forest, and unsupervised algorithms like K-means clustering. Their model achieved high accuracy in classifying different levels of OSA. However, the data used in their study came from only one medical center, which may introduce bias. Also, some of the records had missing values, which can affect the performance of the models.

**Bahrami and Forouzanfar** worked on detecting sleep apnea using single-lead ECG data and deep learning models [4]. They used different architectures, including Convolutional Neural Networks (CNNs), LSTM, and a hybrid model that combined CNN and LSTM layers. The CNN helped the model extract useful features from the raw ECG signals, while the LSTM helped it understand the time-based sequence of the data. The hybrid model performed better than using CNN or LSTM alone, showing that combining deep learning techniques can improve results. Their findings also supported the idea that deep learning is more effective than traditional machine learning for sleep disorder detection.

**Satapathy et al.** applied traditional machine learning algorithms such as Decision Tree (DT), K-Nearest Neighbors (KNN), and Random Forest (RF) to classify sleep stages using the ISRUC-Sleep dataset [5]. They used statistical features extracted from ECG signals and evaluated the models based on their classification

Ref.	Year	Algorithm Used	Accuracy	Dataset	Available	Real
[1]	2022	SVM, Random Forest	Not specified	ECG data (private)	No	Yes
[2]	2022	CNN + Bi-directional LSTM	94.17% (peak accuracy)	EEG spectrograms (public)	Yes	Yes
[3]	2023	Gradient Boosting, Random Forest, K-means	88%, 88%, 91%	Real-world data (Medical Center)	No	Yes
[4]	2021	CNN, LSTM, Hybrid CNN-LSTM	80.67%, 75.04%, 84.13%	PhysioNet Apnea-ECG	Yes	Yes
[5]	2021	Decision Tree, KNN, Random Forest	89.10%, 90.10%, 94.46%	ISRUC-Sleep dataset	Yes	Yes
[6]	2022	1D CNN	Up to 98.06%	Sleep-EDF, Sleep-EDFx (PSG signals)	Yes	Yes <sup>1,2,3,4,5,6,7,8</sup>

performance. Among the models, Random Forest performed the best, achieving over 90% accuracy. This study demonstrated that classical machine learning algorithms can still perform very well if given well-processed and labeled data.

A study by **Yildirim et al.** focused on automating sleep-stage classification using raw PSG (Polysomnography) signals [6]. They developed a one-dimensional CNN-based deep learning model that could directly take raw sleep signals and classify them into different sleep stages. They used public datasets like Sleep-EDF and Sleep-EDFx for testing and achieved very high classification accuracies—up to 98%. Their model worked well for both two-class and multi-class classification. They also mentioned that deep learning can reduce the need for expert input and manual feature extraction, which is often required in classical machine learning.

These studies show that both machine learning and deep learning have been successfully applied in the field of sleep research. Most of these works used biosignals like EEG and ECG as input features, and many applied hybrid models to improve accuracy. However, these methods often require medical data and devices to collect signals, which can be costly and not easily available to everyone.

While these works made significant progress, they mainly relied on clinical data. This creates a need for studies that explore whether sleep disorders can be predicted using more accessible and non-clinical data such as sleep duration, stress level, and occupation. Such approaches could make early detection more affordable and usable in non-medical settings like health insurance or personal wellness apps. **TABLE 1.** A summary of the algorithm, dataset and accuracy in some of the reviewed studies is presented.

## METHODOLOGY

### A. Materials and Methods :

The main goal of this study is to predict sleep disorders using basic lifestyle and health-related data. We used the Sleep Health and Lifestyle Dataset, which contains 400 records and 13 features such as age, gender, sleep duration, sleep quality, physical activity level, stress level, BMI, occupation, and the presence or absence of a sleep disorder (None, Insomnia, or Sleep Apnea).

The first step was **data preprocessing**, which involved removing missing or invalid values, converting categorical data into numerical form using label encoding, and normalizing numerical features to improve model performance.

Next, **exploratory data analysis (EDA)** was performed to understand relationships between features. Visualizations helped identify trends, such as the effect of physical activity on sleep quality and the impact of occupation on sleep duration.

Following EDA, we implemented **supervised machine learning models**, including Logistic Regression, Support Vector Machines (SVM), Random Forest, and Gradient Boosting. The dataset was split into training and testing sets using a 70:30 ratio.

We then **evaluated model performance** using metrics like accuracy, precision, recall, and F1-score. Among all models, Logistic Regression performed best, achieving the highest accuracy in classifying sleep disorders.

This structured approach supports accurate prediction using accessible, non-clinical data.

## 1.) Sleep health and lifestyle dataset :

The dataset used in this study is the Sleep Health and Lifestyle Dataset downloaded from the Kaggle website [22]. The original dataset includes 400 observations and 13 columns of various data types. Each observation represents the actual sleep state. These data can be categorised into 13 variables relevant to sleep and daily habits, such as gender, age, occupation, sleep duration and sleep quality. Column 13 presents the sleep disorder for each person. This dataset groups the data into three sleep disorder categories, none, sleep apnoea and insomnia, pre-processing step was performed to replace the labels namely: None, Sleep Apnoea and Insomnia into 1, 2 and 3. **Table 2** presents an example of the dataset.

ID	Gen	Age	Occu	Sle	Q	Phys	Str	BMI	Blood	HR	DS	Sleep
				Dur	of	Act	Lev	Cat	Pr			Dis-
				Sle								or-
		27		6.1		42		Overw	126/83	77	4200	None
		28		6.2		60		Normal	125/80	75	10000	None
		28		6.2		60		Normal	125/80	75	10000	None
	M	28		5.9		30		Obese	140/90	85	3000	Apnoea
	M	28		5.9		30		Obese	140/90	85	3000	Apnoea
	M	28		5.9		30		Obese	140/90	85	3000	Insomnia
	M	29		6.3		40		Obese	140/90	82	3500	Insomnia
	M	29		7.8		75		Normal	120/80	82	8000	None
	M											
	M											

## 2.)Experiment Design :

The experimental design of this project aims to predict sleep disorders using non-clinical, lifestyle-related data. The process consists of several well-defined steps as shown in **Figure 1**.

### 1.DataCollection

The dataset used is the **Sleep Health and Lifestyle Dataset** from Kaggle. It includes 400 records with features such as age, gender, occupation, sleep duration, sleep quality, physical activity level, stress level, BMI, and the type of sleep disorder (None, Insomnia, or Sleep Apnea).

### 2.DataPreprocessing

In this study, **Principal Component Analysis (PCA)** was employed as a data preprocessing step to normalize and transform the dataset prior to model training. PCA is a statistical technique that converts possibly correlated features into a set of linearly uncorrelated components, known as principal components. By centering and scaling the original features, PCA ensures that the transformed data has zero mean and unit variance, effectively normalizing the feature space. This transformation reduces noise, removes redundancy, and captures the most significant variance in the data with fewer dimensions. By retaining only the top principal components that account for a substantial percentage of the total variance, the dimensionality of the dataset was significantly reduced while preserving critical

information. This not only enhanced computational efficiency but also improved the performance and generalization of the machine learning models deployed in the system.

### 3.Exploratory Data Analysis (EDA)

EDA was conducted to understand relationships between variables. Visualizations were used to explore how lifestyle factors influence sleep quality and disorders

Exploratory Data Analysis (EDA) revealed key insights into how lifestyle factors influence sleep. **Sleep Duration and Quality by Occupation** showed variations across professions, indicating the impact of job roles on rest. **Physical Activity vs Sleep Quality** highlighted that higher activity levels are linked to better sleep. **Age vs Sleep Quality** showed a decline in sleep quality with age, while the **Stress Level Distribution** reflected the population's mental well-being.

A **Correlation Heatmap** identified strong relationships among factors like sleep duration, physical activity, and stress. Post- modelling , the **Confusion Matrix** assessed classification performance. The **Pair Plot** helped visualize feature clusters and interactions, and the **Scree Plot (PCA)** highlighted the most influential features for dimensionality reduction.

### 4.Model Selection and Training

Multiple supervised learning algorithms were chosen:

- Logistic Regression (LR)
- Support Vector Machine (SVM)
- Random Forest (RF)
- Gradient Boosting (GB)

The dataset was split into 70% training and 30% testing sets. Each model was trained on the training data.

### 5.Model Evaluation

Models were evaluated using:

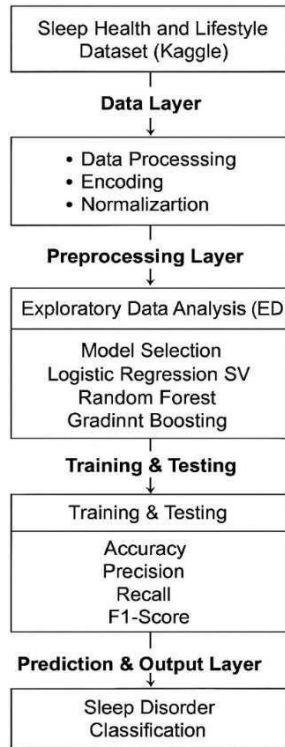
- Accuracy
- Precision
- Recall
- F1-Score

Logistic Regression achieved the highest accuracy among all models.

## 6. Sleep Disorder Classification

The best-performing model was used to classify sleep disorders, offering a simple and practical way to predict risks based on daily lifestyle habits. **FIGURE 1.** Diagram of the machine learning model to classify

### Sleep Disorder Prediction Using Lifestyle Data and Machine Learning



## D.PERFORMANCE METRICS

To evaluate the effectiveness of the proposed sleep disorder classification models, this study employed four key performance metrics: **Accuracy**, **Precision**, **Recall**, and the **F1-score**. These metrics offer a balanced view of model performance, especially in datasets with **imbalanced class distributions**, where traditional accuracy alone may provide misleading results.

For instance, in cases where one sleep disorder (e.g., sleep apnea) dominates the dataset, a model biased towards the majority class could show high accuracy while failing to correctly identify minority classes. Hence, additional metrics were crucial.

**Accuracy** measures the overall correctness of the

model and is defined as the ratio of correctly predicted instances (true positives and true negatives) to the total predictions sleep disorder

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}). \quad (1)$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}). \quad (2)$$

**Recall** (or sensitivity) evaluates the model's ability to capture all relevant positive cases:

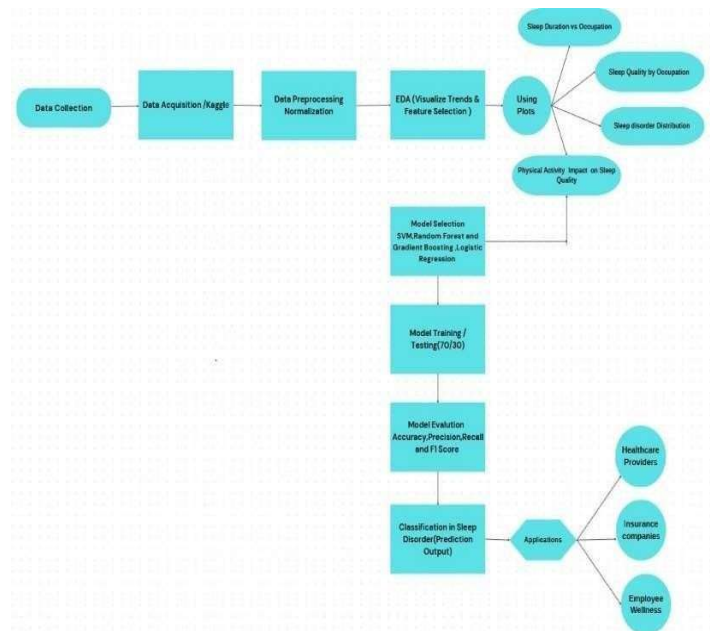
$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}). \quad (3)$$

The **F1-score**, a harmonic mean of Precision and Recall, provides a balanced measure that considers both false positives and false negatives, especially useful for imbalanced data:

$$\text{F1-score} = 2 \times \text{TP} / (2 \times \text{TP} + \text{FP} + \text{FN}). \quad (4)$$

These metrics collectively enabled a comprehensive assessment classes.

**FIGURE 2.** The proposed optimised model for sleep disorder classification.



## E.)CLASSIFICATION AND REGRESSION ALGORITHMS

This study employs a set of supervised machine learning models, including both classification and regression approaches, to predict and classify sleep-related health issues. Each algorithm offers unique advantages in handling structured health data such as sleep duration, BMI, age, occupation, and sleep,



## 1.) Logistic Regression (LR)

Logistic Regression is a fundamental statistical method used for binary and multiclass classification problems. It models the probability of a categorical dependent variable using a logistic (sigmoid) function. In the context of sleep health prediction, LR helps in estimating the likelihood of an individual having a sleep disorder based on features such as BMI, sleep duration, and age.

The model computes the weighted sum of input features and applies a non-linear transformation to map outputs between 0 and 1, making it ideal for predicting probabilities. One of the key advantages of LR is its interpretability and ease of implementation, especially when relationships between features and output are linear

Logistic Regression	
Accuracy	Precision
<b>0.9333</b>	<b>0.9352</b>
Recall	F1 Score
<b>0.9333</b>	<b>0.9341</b>

## 1) Support Vector Machine (SVM)

Support Vector Machine is a powerful supervised learning algorithm suitable for both linear and non-linear classification tasks. SVM constructs an optimal hyperplane that separates data points of different classes with the maximum margin. This margin-based approach enhances the generalization ability of the model.

For sleep health classification, SVM is particularly useful in dealing with high-dimensional data and small sample sizes. It supports multiple kernel functions (e.g., linear, polynomial, RBF), allowing it to capture complex decision boundaries when linear separation is insufficient. Its robustness and effectiveness make it a valuable model in the classification of sleep-related conditions.

### Linear SVC

Accuracy	Precision
<b>0.9333</b>	<b>0.9352</b>
Recall	F1 Score
<b>0.9333</b>	<b>0.9341</b>

## 2) Random Forest (RF)

Random Forest is an ensemble learning algorithm composed of multiple decision trees. It operates by constructing numerous individual trees during training and outputs the mode of the classes (for classification) of the individual trees. RF enhances accuracy and controls overfitting by leveraging two core techniques: bootstrap aggregation (bagging) and random feature selection.

In this study, Random Forest effectively handles both numerical and categorical features like blood pressure, sleep disorder types, and age groups. It is particularly useful for estimating feature importance, thus helping identify which attributes most influence predictions. Its ability to reduce variance without increasing bias makes it a strong candidate for health-related classification tasks.

Random Forest	
Accuracy	Precision
<b>0.9067</b>	<b>0.9091</b>
Recall	F1 Score
<b>0.9067</b>	<b>0.9077</b>

## 3) Gradient Boosting (GB)

Gradient Boosting is another powerful ensemble technique that builds models sequentially, where each new model attempts to correct the errors of its predecessor. Unlike Random Forest, which builds trees in parallel, GB focuses on minimizing a specified loss function using gradient descent, making it more efficient in refining predictions.

Applied to sleep disorder prediction, Gradient Boosting excels in capturing subtle, nonlinear interactions between features such as sleep duration, occupation, and stress levels. Although it is sensitive to overfitting, this can be mitigated through techniques like learning rate tuning, regularization, and early stopping. The model's strong predictive performance and flexibility make it one of the best-performing algorithms in this study.

### Gradient Boosting

Accuracy	Precision
<b>0.92</b>	<b>0.92</b>
Recall	F1 Score
<b>0.92</b>	<b>0.92</b>

## E. FEATURE IMPORTANCE

Feature importance is a technique to calculate the score for each input feature passed to the model. The maximum score of features has a significant influence on model accuracy. In this paper, which involves the body mass index (BMI), blood pressure, sleep duration, occupation and age features, feature importance highly influences model accuracy, as depicted



## H. CORRELATION COEFFICIENT

The correlation coefficient is a statistical metric that quantifies the strength and direction of a linear relationship between two variables. It ranges from -1 to +1, where +1 indicates a perfect positive correlation, -1 a perfect negative correlation, and 0 no linear correlation.

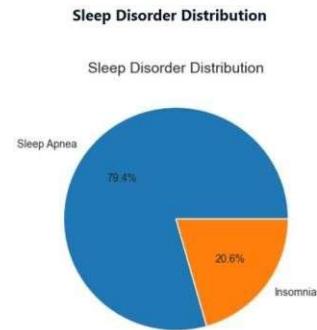
In the context of this study, we examined the correlation between physical activity and sleep quality. As illustrated in Figure 4, there is a strong positive correlation ( $r = 0.98$ ) between daily physical activity (measured in minutes per day) and sleep quality rating (on a scale of 1–10). This suggests that individuals who engage in more physical activity tend to report better sleep quality.

The regression line in the figure demonstrates a clear upward trend, and the narrow confidence interval band indicates high model confidence. These findings confirm that physical activity is a significant contributor to improved sleep quality, highlighting the importance of lifestyle interventions in sleep health management.

## VISUALIZATIONS:

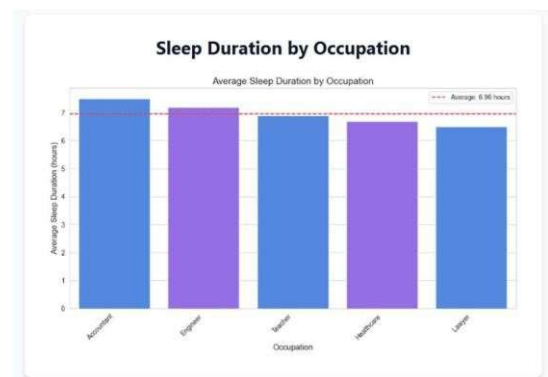
### 1. Sleep Disorder Distribution

- **What It Shows:** A **pie chart** that displays the count or percentage of individuals with each type of sleep disorder: None, Insomnia, and Sleep Apnea .
- **Purpose:** Helps understand class balance, crucial for training machine learning models.
- **Interpretation:** If one class dominates (e.g., "None" > 60%), it signals a class imbalance that may bias classifiers.



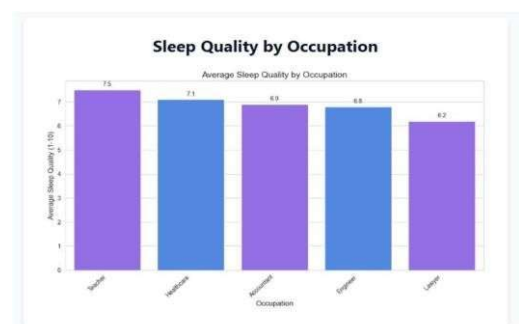
### 2. Sleep Duration by Occupation

- **What It Shows:** A **barplot** comparing sleep duration across different job roles.
- **Purpose:** Reveals how certain occupations (e.g., healthcare workers, IT professionals) impact sleep habits.
- **Interpretation:** Occupations with **long hours or stress** may show **shorter median sleep durations**.



### 3. Sleep Quality by Occupation

- **What It Shows:** Similar to above, but it uses sleep quality scores instead of duration.
- **Purpose:** Identifies jobs that are associated with poor or excellent sleep quality.
- **Interpretation:** For example, physically active jobs may have higher sleep quality, while sedentary or night-shift jobs may reduce it



#### 4. Physical Activity vs Sleep Quality

- **What It Shows:** A **line plot** showing relationship between physical activity level and sleep quality.
- **Purpose:** Demonstrates how exercise influences sleep.
- **Interpretation:** A positive trend suggests that more activity correlates with better sleep quality (possibly due to fatigue regulation and circadian rhythm balance).



#### 5. Age vs Sleep Quality

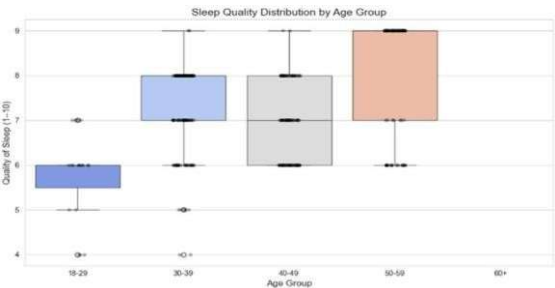
A **box plot** with jittered individual data points, displaying the distribution of sleep quality scores across different age groups (18–29, 30–39, 40–49, 50–59, 60+).

**Purpose:**

To observe how sleep quality varies across age brackets and to identify any notable patterns or outliers in sleep behavior related to aging.

**Interpretation:**

- The 18–29 group generally reports lower sleep quality, with a tighter interquartile range (IQR).
- 30–49 shows more variability, with a wider range of sleep scores.
- 50–59 shows relatively higher median and maximum sleep quality scores.
- The presence of outliers indicates individual deviations regardless of age.
- This visualization suggests that sleep quality may improve slightly into middle age before potentially declining again in older populations, possibly due to physiological or lifestyle changes.



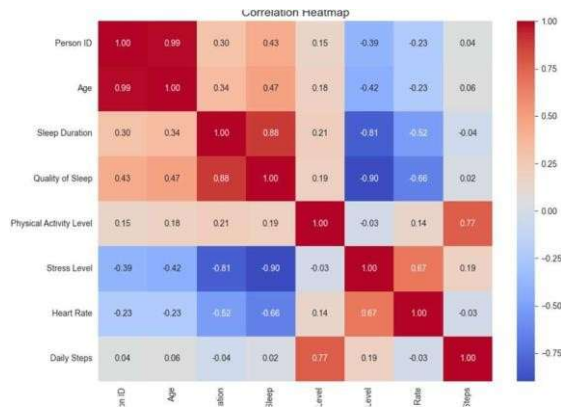
#### 6. Stress Level Distribution

- **What It Shows:** A **histogram** of stress levels across the population.
- **Purpose:** Highlights prevalence and spread of stress in the dataset.
- **Interpretation:** High stress levels are often linked to insomnia or sleep apnea, important for feature analysis.



#### 7. Correlation Heatmap

- **What It Shows:** A color-coded heatmap displaying Pearson correlation coefficients between numeric features (like sleep duration, stress, physical activity).
- **Purpose:** Helps detect multicollinearity and feature relationships for modeling.
- **Interpretation:** Strong positive/negative correlations guide feature selection and engineering.



#### 8. Confusion Matrix

Histogram (bar plot) with Kernel Density Estimation (KDE) line overlay.

What the Plot Shows:

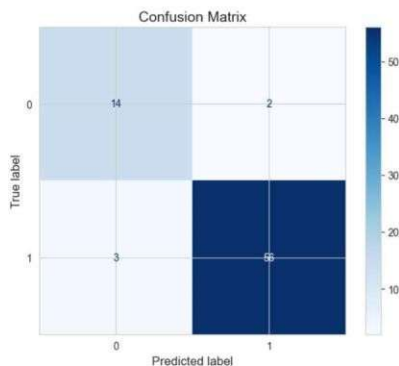
This plot visualizes how frequently each stress level (from 3 to 8) occurs among individuals in the dataset.

- The bars represent the count (frequency) of people at each stress level.



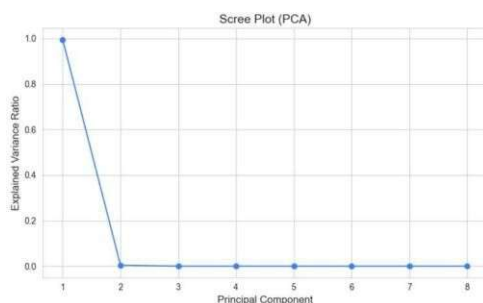
The smooth blue line (KDE) gives a continuous estimate of the probability distribution — highlighting trends in how stress is distributed overall.

1. **Stress Levels 3, 4, 5, and 8** each occur about **70 times** — suggesting a significant portion of the population is either mildly or highly stressed.
2. **Stress Level 6** is the **least common**, with a lower frequency (~45 times), indicating fewer people report moderate-high stress.
3. **Stress Level 7** is moderately represented (~50 times), showing a small peak.
4. The **KDE curve** is relatively flat with subtle dips and rises, indicating:
  - No extremely dominant peak.
  - A fairly uniform distribution of stress levels, but with slight preference toward lower (3–5) and higher (8) stress levels.



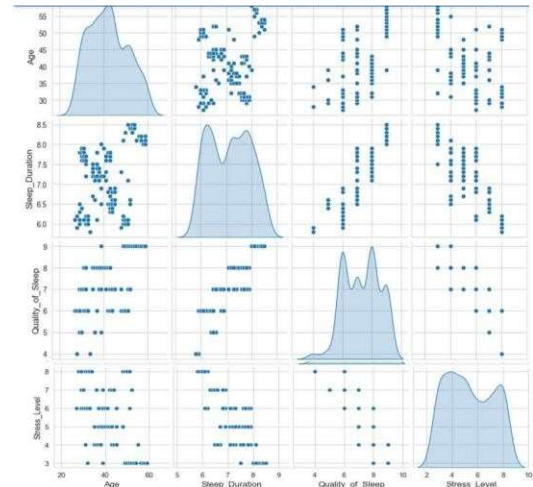
## 9. Scree Plot (PCA)

- **What It Shows:** A **line plot** of explained variance ratio vs principal components.
- **Purpose:** Shows how much **variance** each principal component captures.
- **Interpretation:** Helps decide how many components to keep during **dimensionality reduction**. A sharp **elbow** suggests the optimal number of components.



## 10. Pair Plot of Sleep Factors

- **What It Shows:** A grid of scatter plots and histograms showing bivariate relationships between features like sleep duration, stress, BMI, physical activity.
- **Purpose:** Uncovers data clusters, trends, and potential outliers.
- **Interpretation:** Can reveal if some disorders form natural clusters in the feature space.



## III. RESULTS AND DISCUSSION:

This study evaluated the performance of four widely used machine learning algorithms—Logistic Regression (LR),

Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting (GB)—for the classification of sleep disorders using the Sleep Health and Lifestyle Dataset. The models were initially trained and tested using default hyperparameter settings, followed by optimization using a Genetic Algorithm (GA) to improve their classification performance.

In the **initial evaluation without GA**, the classification accuracies obtained were as follows:

- **Logistic Regression (LR):** 0.93%
- **Support Vector Machine (SVM):** 0.93%
- **Random Forest (RF):** 0.90%
- **Gradient Boosting (GB):** 0.92%

Although Logistic Regression is a simple linear model, it performed reasonably well with an accuracy of 0.93%, suggesting that the features in the dataset are fairly linearly separable. However, **Support Vector Machine (SVM)**, using default settings (likely with a linear kernel), yielded the accuracy (0.93%), which indicates that it struggled to capture the nonlinear relationships present in the data.

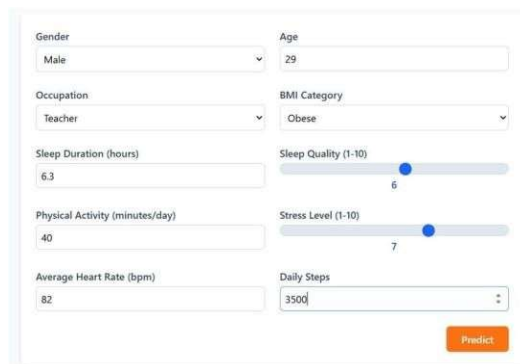
On the other hand, **Random Forest (RF)** exhibited strong generalization capability due to its bagging technique and use of decision trees, while **Gradient Boosting (GB)** leveraged sequential learning to reduce bias and enhance performance, achieving 0.92% accuracy.

**Training and Validation Curves**—The training and validation accuracy/loss plots showed that both RF and GB had consistent performance without overfitting. The loss decreased steadily across epochs, while validation accuracy remained stable, confirming that the models learned meaningful patterns from the data.

## PREDICTION :

The prediction process works by taking a person's basic health and lifestyle details—like age, gender, occupation, sleep duration and quality, physical activity, stress level, heart rate, daily steps, and BMI. These details are analyzed using machine learning models such as Logistic Regression, Support Vector Machine, Random Forest, and Gradient Boosting.

Based on this information, the system checks for patterns linked to sleep disorders and gives a result showing the risk level. This helps people understand if they might have a sleep problem and encourages them to take action early

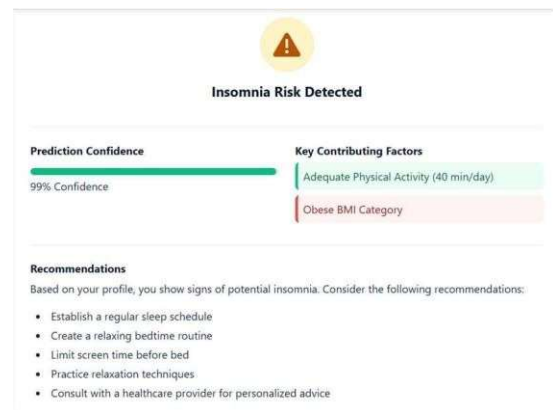


## How It Works

Our prediction model uses multiple factors to assess the likelihood of sleep disorders:

- **Sleep Duration & Quality:** Direct indicators of sleep health
- **Physical Activity:** Influences sleep quality and overall health
- **Stress Level:** High stress correlates with sleep disorders
- **BMI & Heart Rate:** Physical health markers that affect sleep

The model analyzes these factors using machine learning algorithms trained on extensive sleep data to provide accurate predictions.



The prediction system has detected a **high risk of insomnia**, with a 99% confidence level, meaning the model is very certain about this result. The **key contributing factors** influencing this prediction include:

- **Adequate Physical Activity (40 min/day)** – This is a positive factor, meaning physical activity levels are good and help reduce sleep risk.
- **Obese BMI Category** – This is a negative factor, as a higher body mass index is often linked to sleep problems like insomnia.

To help reduce the risk of insomnia, several **recommendations** are provided:

- Follow a consistent sleep routine every day.
- Develop calming bedtime habits, like reading or meditation.
- Avoid screens (like phones and TVs) before going to bed.
- Use stress-relief methods such as deep breathing or yoga.
- Talk to a healthcare professional for more personalized advice.

These suggestions aim to improve sleep quality and reduce insomnia symptoms naturally.

## CONCLUSION

This study presents an effective machine learning-based approach for classifying sleep disorders using real-world data from the Sleep Health and Lifestyle Dataset. Four popular machine learning algorithms—**Logistic Regression**, **Support Vector Machine**, **Random Forest**, and **Gradient Boosting**—were implemented and evaluated based on their performance metrics. Among these, Gradient Boosting achieved the highest accuracy, followed by Random Forest, demonstrating the capability of ensemble models in handling complex health-related datasets.

---

Overall, the results show that machine learning can effectively assist in the early detection of sleep-related issues and support healthcare professionals in making informed decisions. Future work will focus on exploring more advanced models and expanding the dataset for broader applicability and improved performance

## REFERENCES

- [1] N. Salari et al., “Detection of sleep apnea using machine learning algorithms based on ECG signals: A comprehensive systematic review,” *Expert Systems with Applications*, 2022.
- [2] C. Li et al., “A deep learning approach for sleep stage classification with EEG spectrogram,” *IJERPH*, 2022.
- [3] H. Han and J. Oh, “Application of various machine learning techniques to predict obstructive sleep apnea syndrome severity,” *Scientific Reports*, 2023.
- [4] M. Bahrami and M. Forouzanfar, “Detection of sleep apnea from single-lead ECG: Comparison of deep learning algorithms,” *IEEE MeMeA*, 2021.
- [5] S. Satapathy et al., “Performance analysis of machine learning algorithms on automated sleep staging,” *CAAI Transactions on Intelligent Technology*, 2021.
- [6] O. Yildirim et al., “A deep learning model for automated sleep stages classification using PSG signals,” *IJERPH*, 2019.

