SOMNISENSE

A Data Driven ML system for sleep disorder classificationn

Submitted By:

K Subiksh sundar

Mohammed Zunaid Ahmed

Aarju Shaw

Sai Pujitha Velpuri

Problem Overview

Sleep disorders such as Insomnia and Sleep Apnea have become increasingly prevalent due to:

- High stress levels
- Sedentary lifestyle patterns
- Poor sleep hygiene
- Cardiovascular and metabolic risk factors

Project Objective

The primary objective of this project is to develop a machine learning-based prediction system that:

- Analyzes demographic, lifestyle, and physiological indicators
- Identifies key factors influencing sleep disorders
- Classifies individuals into risk categories (None, Insomnia, Sleep Apnea)
- Provides analytical insights through interactive dashboards
- Demonstrates real-time prediction using a Streamlit web application

DATASET OVERVIEW & FEATURE ENGINEERING

Dataset Description

The dataset contains structured demographic, lifestyle, and physiological information collected to analyze patterns associated with sleep disorders.

- Total Records: 374
- Total Features: 14
- Classification Type: Multi-class (None, Insomnia, Sleep Apnea)
- Data Type: Structured tabular dataset

The dataset integrates behavioral, cardiovascular, and sleep-related indicators to enable predictive modeling.
Feature Categories

Feature Categories

Demographic Attributes

- Age
- Gender
- Occupation

These variables help assess population-level sleep risk patterns.

Lifestyle & Behavioral Indicators

- Physical Activity Level
- Daily Steps
- Stress Level

These factors capture behavioral influences affecting sleep health.

Sleep & Physiological Metrics

- Sleep Duration
- Quality of Sleep
- BMI Category
- Blood Pressure
- Heart Rate

These features represent biological and cardiovascular health indicators associated with sleep

Exploratory Data Analysis (EDA)

Objective of Analysis

To systematically examine patterns, relationships, and statistical trends within demographic, lifestyle, and physiological variables in order to identify key factors influencing sleep disorders and support predictive modeling.

Class Distribution Analysis

The dataset contains three categories: None, Insomnia, and Sleep Apnea.

Insomnia cases are comparatively higher than Sleep Apnea cases.

The class distribution indicates meaningful variation suitable for multi-class classification.

Behavioral & Lifestyle Insights

Higher stress levels show a strong positive correlation with Insomnia.

Individuals with lower physical activity and fewer daily steps demonstrate increased disorder probability.

Reduced sleep duration significantly contributes to sleep imbalance.

Poor sleep quality is one of the strongest indicators of disorder classification.

Physiological & Health Indicators

Overweight and obese BMI categories show higher prevalence of Sleep Apnea.

Elevated systolic and diastolic blood pressure values are associated with increased disorder risk.

Heart rate variations demonstrate moderate correlation with sleep abnormalities.

Age shows gradual increase in sleep disorder probability, particularly in middle-aged individuals.

Outcome of EDA

- Identified significant predictors for model training.
- Improved feature selection strategy.
- Strengthened understanding of health-risk relationships.
- Provided analytical foundation for machine learning development.

MACHINE LEARNING MODEL DEVELOPMENT

Problem Formulation

The task was defined as a multi-class classification problem to predict the type of sleep disorder based on demographic, lifestyle, and physiological features.

Target Classes:

- No Sleep Disorder
- Insomnia
- Sleep Apnea

The dataset was structured for supervised learning using labeled health data.

Model Implementation

Two classification algorithms were developed and compared:

- Decision Tree Classifier – Used as a baseline model due to its interpretability and ability to capture non-linear relationships.
- Random Forest Classifier – Ensemble-based model that improves predictive accuracy by combining multiple decision trees.

Feature engineering steps included:

- Encoding categorical variables
- Splitting Blood Pressure into Systolic and Diastolic values
- Selecting significant predictors identified during EDA

Model Evaluation & Final Selection

Model performance was assessed using:

- Accuracy
- Precision
- Recall
- F1-Score

The Random Forest model demonstrated better generalization, improved class-wise performance, and reduced overfitting compared to the Decision Tree model.

Therefore, Random Forest was selected for deployment in the Streamlit application.

SYSTEM IMPLEMENTATION & DEPLOYMENT

Power BI Dashboard

An interactive analytical dashboard was designed to translate data insights into visual interpretations for better decision-making.

The dashboard provides:

Distribution of sleep disorder categories

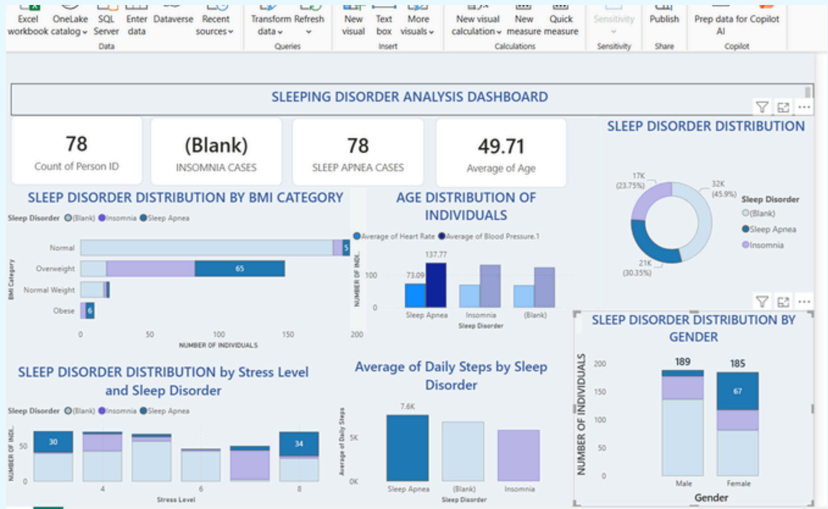
Gender-wise and age-wise comparison analysis

BMI category impact on Sleep Apnea prevalence

Stress level and sleep duration influence visualization

Cardiovascular indicators (Blood Pressure & Heart Rate) correlation

The dashboard enables intuitive understanding of patterns discovered during EDA and supports analytical storytelling.



Streamlit Web Application

A dynamic web application was developed using Streamlit to demonstrate real-time prediction capability.

The application allows users to:

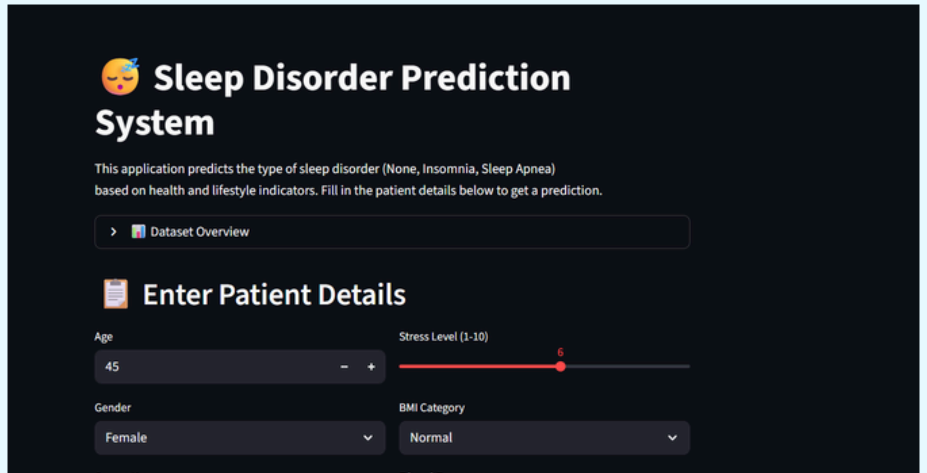
Enter demographic and lifestyle details

Provide health-related inputs (BMI, stress level, sleep duration, etc.)

Receive instant model-based prediction output

Interact with a clean and structured user interface

This component transforms the machine learning model into a deployable and user-accessible solution.



😴 Sleep Disorder Prediction System

This application predicts the type of sleep disorder (None, Insomnia, Sleep Apnea) based on health and lifestyle indicators. Fill in the patient details below to get a prediction.

› 📊 Dataset Overview

📋 Enter Patient Details

Age: 45

Stress Level (1-10): 6

Gender: Female

BMI Category: Normal

Backend Implementation

The backend integrates machine learning logic with deployment infrastructure.

Key backend processes include:

Loading the trained Random Forest model

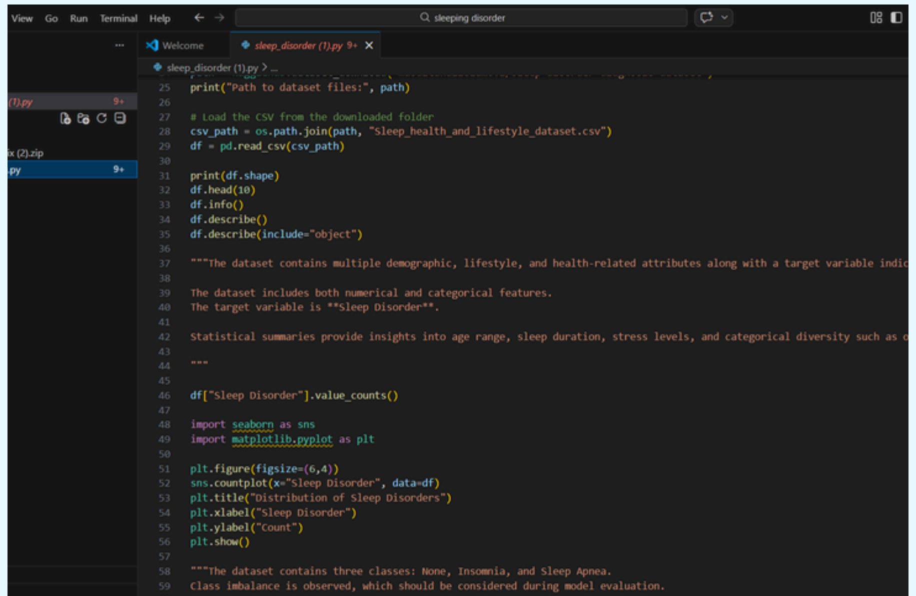
Applying feature transformations and encoding

Processing user inputs dynamically

Generating prediction outputs

Ensuring smooth integration between model and UI

The backend ensures reliability, scalability, and structured model execution within the web application



```
print("Path to dataset files:", path)

# Load the CSV from the downloaded folder
csv_path = os.path.join(path, "Sleep_health_and_lifestyle_dataset.csv")
df = pd.read_csv(csv_path)

print(df.shape)
df.head(10)
df.info()
df.describe()
df.describe(include="object")

"""The dataset contains multiple demographic, lifestyle, and health-related attributes along with a target variable indic

The dataset includes both numerical and categorical features.
The target variable is **Sleep Disorder**.

Statistical summaries provide insights into age range, sleep duration, stress levels, and categorical diversity such as o

"""

df["Sleep Disorder"].value_counts()

import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(6,4))
sns.countplot(x="Sleep Disorder", data=df)
plt.title("Distribution of Sleep Disorders")
plt.xlabel("Sleep Disorder")
plt.ylabel("Count")
plt.show()

"""The dataset contains three classes: None, Insomnia, and Sleep Apnea.
Class imbalance is observed, which should be considered during model evaluation.
```

CONCLUSION & FUTURE SCOPE

Conclusion

This project successfully developed a data-driven machine learning system to predict sleep disorders using demographic, lifestyle, and physiological factors.

Key outcomes:

- Identified significant predictors influencing sleep health
- Developed and evaluated multiple classification models
- Selected Random Forest for optimal predictive performance
- Designed an interactive Power BI dashboard for analytical insights
- Deployed a functional Streamlit web application for real-time prediction

The system demonstrates how data science techniques can support healthcare risk assessment and early detection strategies.

Future Scope

The system can be further enhanced by:

- Integrating real-time wearable device data
- Expanding the dataset for improved generalization
- Incorporating deep learning models for higher accuracy
- Deploying on cloud infrastructure for large-scale access
- Adding automated health recommendations based on predictions