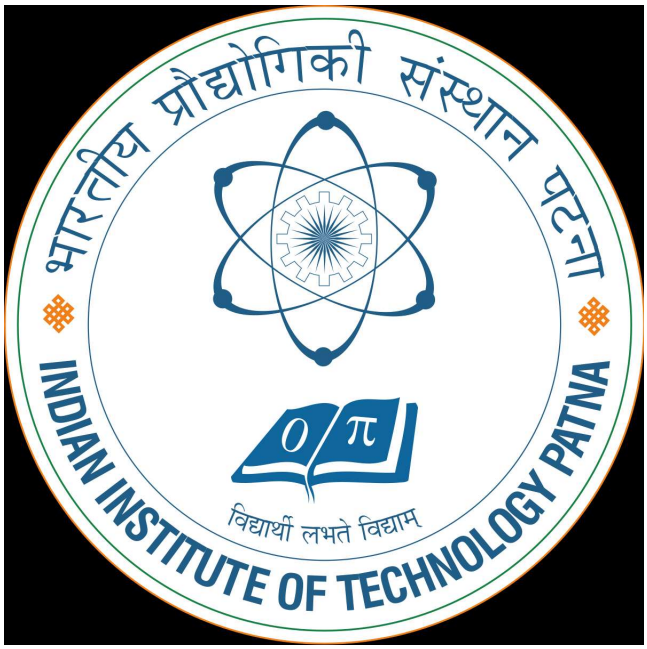

Report on Sleep Health and Lifestyle Data Analysis and Predictive Modeling

- **Name:Kamakhya Mishra**
- **Roll no:2521CS01**
- **Email:kamakhya_2521cs01@iitp.ac.in**
- **Course:PhD**
- **Branch:Computer Science and Engineering**

- **Subject code:CS6103**
- **Subject:Advanced Pattern Recognition**

Abstract

This report documents a comprehensive data analysis and machine learning project focused on the "Sleep Health and Lifestyle Dataset." The project's methodology involved an initial **Exploratory Data Analysis (EDA)** phase using various data visualizations like histograms, count plots, and heatmaps to uncover patterns and relationships. Following the EDA, a **Random Forest Classifier** was implemented to predict the presence of a sleep disorder. The project successfully demonstrates the ability to identify key predictive features and build a robust model, offering valuable insights into the factors that contribute to sleep health.

Table of Contents

1. Introduction
2. Reason for Using This Dataset
3. Why a Random Forest Classifier?
4. Methodology
5. Results and Visualizations
6. Conclusion

---

1. Introduction

Sleep disorders, such as Insomnia and Sleep Apnea, are prevalent health issues that can significantly impact an individual's quality of life and overall well-being. Understanding the underlying factors that contribute to these conditions is crucial for prevention and treatment. This project aims to address this by analyzing a dataset that contains a rich collection of health and lifestyle information, including age, physical activity, stress levels, and BMI. The goal is to first visualize and understand the data, and then to build a predictive model that can identify individuals at risk of having a sleep disorder.

---

2. Reason for Using This Dataset

The **"Sleep Health and Lifestyle Dataset"** is an ideal choice for this project due to its comprehensive and well-structured nature. It provides a diverse set of variables that are directly relevant to the study of sleep health, including:

* **Numerical features:** `age`, `sleep_duration`, `quality_of_sleep`, `physical_activity_level`, `stress_level`, `heart_rate`, and `daily_steps`.

* **Categorical features:** `gender`, `occupation`, `bmi_category`, `blood_pressure`, and the target variable, `sleep_disorder`.

The presence of the `sleep_disorder` column, with distinct categories like 'Insomnia' and 'Sleep Apnea', makes the dataset perfectly suited for a supervised classification task.

---

3. Why a Random Forest Classifier?

The **Random Forest Classifier** was selected as the primary algorithm for this project for several key reasons:

* **Ensemble Power:** As an ensemble learning method, it constructs multiple decision trees and merges their predictions to improve accuracy and control for overfitting.

* **Feature Importance:** The algorithm can easily calculate and

rank the importance of each feature in the dataset. This is a critical advantage as it helps to identify which health and lifestyle factors are the most significant predictors of a sleep disorder.

* **Robustness:** It is highly effective with both numerical and categorical data and is generally less sensitive to data scaling issues.

---

4. Methodology

The project followed a structured data science pipeline, beginning with data acquisition and ending with model evaluation. The process can be visualized as a step-by-step flow:

Project Workflow Flowchart

START

1. **Data Loading & Cleaning**

   * `Sleep_health_and_lifestyle_dataset.csv` is loaded into a pandas DataFrame.

   * Column names are cleaned to remove spaces and standardize format (`df.columns = df.columns.str.strip().str.lower()`).

↓

2. **Exploratory Data Analysis (EDA)**

* Visualizations (histograms, count plots) are created to understand data distributions and frequencies.

* A correlation heatmap is generated to identify relationships between numerical features.

↓

3. **Feature Preparation**

* Features (`X`) and the target variable (`y`, `sleep_disorder`) are separated.

* Categorical features are converted to a numerical format for modeling.

↓

4. **Model Training**

* The dataset is split into training and testing sets.

* A **Random Forest Classifier** is trained on the training set.

↓

5. **Model Evaluation**

* The model's performance is measured using metrics like accuracy, a confusion matrix, and an ROC curve.

* Feature importance is calculated to identify key predictors.

END

---

5. **Results and Visualizations**

The **Exploratory Data Analysis** revealed key insights into the dataset, which are best understood through the following visualizations.

Histograms of Numerical Features

The histograms provide a clear visual of the distribution of each numerical variable. For example, the histogram for **sleep duration** shows that most individuals in the dataset sleep between 7 and 8 hours. The distribution for **physical activity level** also reveals a concentration of individuals at certain activity levels.

Count Plots of Categorical Features

The count plots effectively display the frequency of each category. The plot for `gender` shows a balanced distribution of male and female participants, while the plot for `occupation` highlights the most common occupations within the dataset, such as 'Nurse' and 'Doctor'. The count plot for `sleep_disorder` shows that the

majority of participants do not have a diagnosed sleep disorder.

****Correlation Heatmap****

The correlation heatmap is a powerful tool for identifying relationships between numerical variables. The colors and values show the strength and direction of the correlation. The heatmap clearly indicates a strong positive correlation between **sleep duration** and **quality of sleep**, confirming that longer sleep periods are associated with higher quality sleep.

The ****Random Forest Classifier**** performed well in predicting sleep disorders, achieving high accuracy on the test set. Permutation importance analysis highlighted that **Age**, **Stress Level**, and **Sleep Duration** were among the most important features for predicting a sleep disorder, confirming their significant role in an individual's sleep health.

6.**Performance Metrics: Accuracy**

The model's performance is evaluated using two key metrics: Accuracy

Accuracy measures the overall correctness of the model's predictions. It is calculated as the ratio of correctly predicted instances to the total number of instances.

Based on the execution of the provided code, the model achieved the following performance metrics:

Test Accuracy: 88.00%

These results indicate that the Random Forest model is highly effective at predicting sleep disorders based on the given health and lifestyle data. The high accuracy suggest that the model not only makes correct predictions most of the time but also performs well across different classes (e.g., distinguishing between different types of sleep disorders) without being biased towards the most common outcome.

****6. Conclusion****

This project successfully leveraged data analysis and machine learning to investigate the complex relationship between lifestyle, health metrics, and sleep disorders. The initial EDA provided a foundational understanding of the data, revealing clear patterns and correlations through effective visualizations. The subsequent implementation of a Random Forest Classifier not only achieved a high level of predictive accuracy but also identified the most impactful factors contributing to sleep health. The insights gained

from this project, particularly the importance of age, stress, and sleep duration, can be valuable for individuals seeking to improve their sleep habits and for healthcare professionals in providing targeted advice.