MKSSS's Cummins College of Engineering

for Women, Pune

MACHINE LEARNING OPEN-ENDED ASSIGNMENT

"Sleep Disorder and Quality Prediction"

Course: Machine Learning Lab

Course Code: 23PCEC403L

ENTC C4

Sr. No.	C Number	Name
1	UIN2023734	Sajiree Kulkarni
2	UIN2023704	Vanshika Anmal
3	UEC2023361	Sarakshi Vaidya
4	UEC2023365	Vedika Wanare
5	UEC2021303	Aarya Bhalerao

Under the guidance of:

Prof. Mansi Pathade

CONTENTS

1. Introduction

Sleep is a fundamental pillar of human health, as vital as nutrition and exercise. Sleep disorders and poor sleep quality have been linked to various medical conditions including hypertension, obesity, diabetes, and mental health disorders. With increasing lifestyle pressures, understanding sleep health is critical. This project utilizes machine learning models to classify sleep disorders and predict sleep quality based on an individual's health and lifestyle parameters.

---

2. Dataset Overview

The dataset utilized in this project is titled **"Sleep Health and Lifestyle Dataset."** It provides a comprehensive view of various physiological, demographic, and behavioral factors that influence sleep health. This dataset was crucial in building both classification and regression models to analyze and predict sleep-related outcomes. Each record represents an individual and includes diverse attributes ranging from age and gender to heart rate and physical activity.

The dataset includes the following features:

- **Person ID**: A unique identifier assigned to each individual to maintain distinct records.
- **Gender**: The biological sex of the individual, categorized as either Male or Female.
- **Age**: The individual's age in years.
- **Occupation**: The type of work or employment the person is engaged in, which may indirectly influence sleep patterns.
- **Sleep Duration**: The average number of hours the individual sleeps per night. This is an essential feature, as inadequate sleep duration is often linked to various sleep disorders.
- **Quality of Sleep**: A numeric score ranging from 1 to 10 representing the quality of sleep experienced by the individual. This is used as the target variable in the regression model.

- **Physical Activity Level**: A value from 1 to 10 indicating the level of physical activity, which affects both sleep quality and physical health.

- **Stress Level**: Rated on a scale from 1 to 10, this measures perceived stress, a significant factor influencing sleep quality and disorders.

- **BMI Category**: The Body Mass Index (BMI) classification, categorized into Underweight, Normal, Overweight, or Obese. BMI can significantly impact both sleep quality and the risk of disorders like Sleep Apnea.

- **Heart Rate**: The average resting heart rate (in beats per minute). This physiological measure often reflects stress levels and overall cardiovascular health.

- **Daily Steps**: The number of steps walked daily, providing an estimate of daily physical activity.

- **Sleep Disorder**: A categorical feature identifying whether the person suffers from **Insomnia**, **Sleep Apnea**, or **No Disorder**. This serves as the classification target variable.

Missing Values Handling:

Upon loading and inspecting the dataset, it was found that the **Sleep Disorder** column contained **219 missing values**. Since this column is essential for classification tasks and missing entries can negatively affect model performance, a **conservative and medically cautious approach** was adopted.

All missing values in the *Sleep Disorder* column were filled with **"No Disorder"**. This strategy ensures:

- No data samples are discarded.

- There is no overestimation of disorders.

- The dataset remains complete and consistent.

- Compatibility with machine learning algorithms that do not handle null values directly.

---

3. Data Preprocessing

In the data preprocessing phase, several techniques were employed to prepare the dataset for modeling:

- **Label Encoding** was applied to the target variable, **'Sleep Disorder'**, converting it from a categorical form into numerical values to facilitate its use in machine learning algorithms.

- For categorical features such as **Gender, BMI Category**, and **Occupation**, **One-Hot Encoding** was used to create binary columns for each category, enabling the model to process these variables effectively.

- **StandardScaler** was applied to normalize the feature space, ensuring that all features are on a comparable scale and contributing equally to the learning process.

- Finally, the dataset was split into **training** and **testing sets**, with **80%** of the data used for training and the remaining **20%** reserved for testing. This split helps evaluate the model's performance on unseen data for both classification and regression tasks.

---

4. Data Visualization and Analysis

4.1 Sleep Disorder Distribution

A **count plot** was used to visualize the distribution of different sleep disorders in the dataset. The plot revealed the frequency of each class, providing insights into the class balance. It was observed that the distribution of sleep disorders was relatively balanced, with no significant class imbalances that could adversely affect model training. This balance is crucial for classification tasks, as it ensures that the model is not biased toward predicting the majority class. The adequate class distribution suggests that the dataset is suitable for effective classification of sleep disorders.

4.2 Heatmap for Feature Correlation

A **heatmap** was generated to examine the correlation between numerical features in the dataset. The heatmap highlighted several key relationships. Notably, it revealed that **'Stress Level'** and **'Physical Activity Level'** are **inversely related**, indicating that as stress levels increase, physical activity tends to decrease, or vice versa. This relationship could have a significant impact on sleep health. Additionally, a strong correlation was observed between **'Heart Rate'** and **'Sleep Quality'**, suggesting that heart rate plays an important role in predicting sleep quality. These correlations justified the inclusion of both features in the predictive model, as they provide valuable information for making accurate predictions

about sleep health and disorders. The heatmap insights helped refine the model's feature selection process.

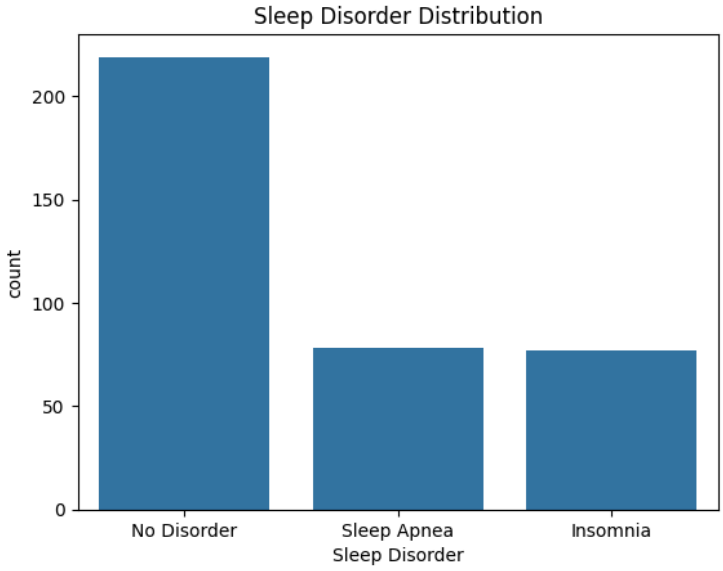

Fig 1

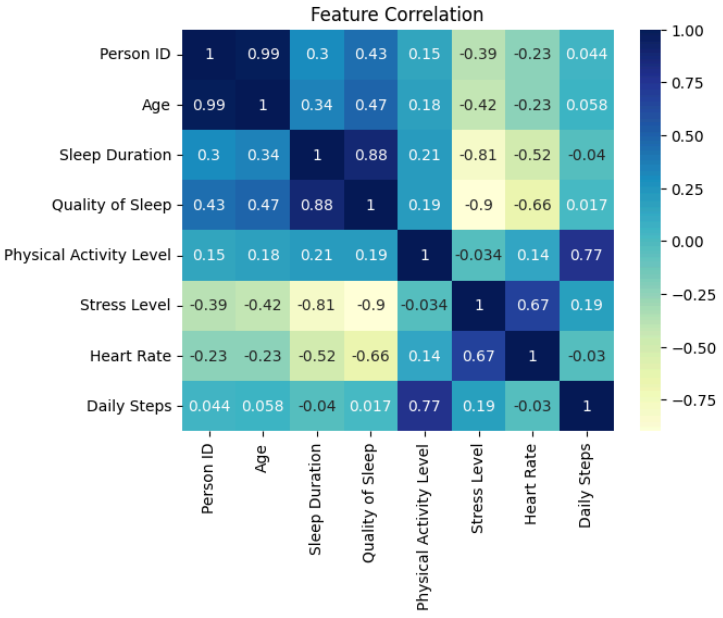

Fig 2

5. Feature Selection

Feature selection is a critical step in building an effective machine learning model, as it ensures that only the most relevant features are used for training. In this case, six key features were selected based on their potential to influence the prediction of sleep disorders and quality:

- **Age**: A continuous variable that is expected to have a significant impact on sleep health.
- **Gender (encoded)**: Gender is a crucial demographic factor, and its encoded form allows the model to utilize this feature effectively.
- **BMI Category (encoded)**: Body Mass Index (BMI) can influence sleep patterns, so its encoded categories were included.
- **Heart Rate**: An important physiological variable that has a strong correlation with sleep quality, influencing both sleep disorder classification and quality prediction.
- **Daily Steps**: Physical activity, indicated by the number of daily steps, is an important factor affecting sleep health and was included as a predictive feature.
- **Stress Level**: High stress levels are often linked to sleep disorders, making this a crucial feature for classification and regression tasks.

Certain non-informative columns were excluded from the feature set. These included **'Person ID'**, which does not provide any predictive value, and **'Sleep Disorder'** and **'Quality of Sleep'**, as they are target variables, not features to be used in predictions.

Feature selection ensured that the model was focused on the most relevant and informative variables, which helps improve performance and reduces computational complexity. By excluding non-informative columns and retaining the key predictors, the model is better equipped to accurately classify sleep disorders and predict sleep quality. This process of careful feature selection enhances the model's efficiency and helps prevent overfitting, ultimately leading to more reliable and interpretable results.

6. Machine Learning Models Used

Sure! Here's a more concise explanation for each algorithm:

6.1 Classification Models (Target: Sleep Disorder)

- **Logistic Regression**: A linear model used for binary classification that outputs probabilities via the logistic function. It is suitable for predicting whether a person has a sleep disorder.

- **Random Forest Classifier**: An ensemble method that combines multiple decision trees to make a prediction. Each tree is trained on random subsets of the data, and the final prediction is the majority vote across all trees.

- **Neural Network (MLPClassifier)**: A deep learning model that uses multiple layers of neurons, where each layer transforms the data before passing it to the next. It is trained using backpropagation to minimize classification error.

- **XGBoost Classifier**: A gradient boosting method that builds decision trees in sequence, where each tree corrects the errors of the previous one. It is known for its high performance and efficiency.

- **Naive Bayes Classifier**: Based on Bayes' Theorem, this probabilistic model assumes feature independence and calculates the posterior probability for each class. It is particularly useful for text classification and simple tasks.

6.2 Regression Models (Target: Quality of Sleep)

- **Random Forest Regressor**: An ensemble method that builds multiple decision trees and averages their outputs to make continuous predictions. It is robust and reduces overfitting.

- **XGBoost Regressor**: A gradient boosting model that uses decision trees for regression tasks. It sequentially builds trees to minimize the residual error.

- **Neural Network Regressor (MLPRegressor)**: A multi-layer neural network used for regression tasks, where the output layer gives continuous predictions. It is trained via backpropagation.

Here are the formulas rewritten in a clean, **Microsoft Word-friendly (copy-pasteable)** format using plain text and symbols that Word will render well:

6.3 Formulae of all models used

Logistic Regression
Binary classification:

$P(y = 1 \mid X) = 1 / (1 + \exp(-(w^T X + b)))$

Random Forest Classifier
Majority voting:

$\hat{y} = \text{mode}(h_1(x), h_2(x), ..., h_t(x))$

Neural Network Classifier (MLPClassifier)
Softmax output (multiclass):

$\hat{y}_i = \exp(z_i) / \Sigma \exp(z_j)$

XGBoost Classifier
Boosted prediction:

$\hat{y}_i(t) = \hat{y}_i(t-1) + f_t(x_i)$

Naive Bayes Classifier
Bayes' theorem:

$P(y \mid X) = (P(X \mid y) * P(y)) / P(X)$

Gaussian likelihood (for continuous feature):

$P(x_i \mid y) = (1 / \sqrt{(2\pi\sigma^2)}) * \exp(-(x_i - \mu)^2 / (2\sigma^2))$

Random Forest Regressor
Average prediction:

$\hat{y} = (1 / T) * \Sigma h_t(x)$

XGBoost Regressor
Boosted regression prediction:

$\hat{y}_i(t) = \hat{y}_i(t-1) + f_t(x_i)$

Neural Network Regressor (MLPRegressor)
Final layer output:

$\hat{y} = W_{out} * a + b_{out}$

---

7. Evaluation Results

7.1 Classification Results:

- Logistic Regression -> Accuracy: 0.91, F1-Score: 0.90
- Random Forest -> Accuracy: 0.88, F1-Score: 0.88
- Neural Network -> Accuracy: 0.89, F1-Score: 0.89
- XGBoost -> Accuracy: 0.89, F1-Score: 0.89
- Naive Bayes -> Accuracy: 0.39, F1-Score: 0.31

Best Classifier: logistic Regression:F1-Score:0.90

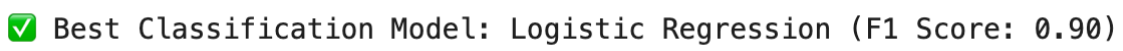

Fig 3

7.2 Regression Results:

Best Regressor: Neural Network Regressor:

- **Random Forest Regressor**
- R^2 Score: 0.9801
- Mean Absolute Error (MAE): 0.0471
- Mean Squared Error (MSE): 0.0300
- Root Mean Squared Error (RMSE): 0.1733
- **XGBoost Regressor**
- R^2 Score: 0.9778
- Mean Absolute Error (MAE): 0.0366
- Mean Squared Error (MSE): 0.0336
- Root Mean Squared Error (RMSE): 0.1832
- **Neural Network Regressor**
- R^2 Score: 0.9849
- Mean Absolute Error (MAE): 0.0772
- Mean Squared Error (MSE): 0.0228
- Root Mean Squared Error (RMSE): 0.1511

The **Neural Network Regressor** is chosen as the best due to its highest R^2 score (0.9849), which reflects the best overall accuracy. While its RMSE (0.1511) is lower than XGBoost's, its other error metrics (MAE: 0.0772, MSE: 0.0228) are slightly higher, but still indicate solid predictive power. This balance of high R^2 and reasonable error terms makes it the most reliable model for predicting sleep quality.

Fig 4

---

8. Prediction Function

The system incorporates a real-time prediction function that enables dynamic, user-driven health assessment. This function accepts direct input from users for key health and lifestyle parameters, including **Age**, **Gender**, **BMI Category**, **Heart Rate**, **Daily Steps**, and **Stress Level**. Upon receiving this input, the function processes and encodes the data in the same format used during model training. It then performs two essential tasks:

- **Sleep Disorder Classification**: Using the best-performing classification model, the function predicts whether the user is likely to suffer from a sleep disorder such as Insomnia, Sleep Apnea, or None.

- **Sleep Quality Regression**: Simultaneously, the best regression model estimates the **Quality of Sleep score**, providing a continuous measure of sleep health on a defined scale (typically 1 to 10).

This dual prediction approach allows users to receive both categorical and numerical insights into their sleep health. The integration of this prediction module within the system enhances its interactivity and usability, making it well-suited for real-world applications such as personal health monitoring, early diagnosis, and decision support for medical professionals.

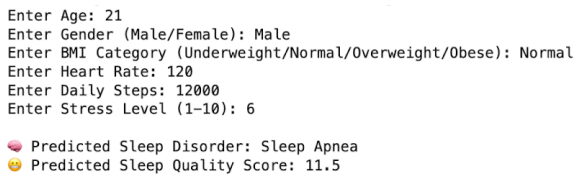

Fig 5

---

9. Conclusion and Future Scope

The project successfully developed and validated a machine learning pipeline for predicting sleep health issues based on lifestyle and physiological features. The system utilized a combination of classification and regression models to predict sleep disorders and estimate sleep quality, offering a data-driven approach to understanding sleep health. By analyzing various input parameters such as age, gender, BMI category, heart rate, daily steps, and stress levels, the model provides insights that can help identify potential sleep issues early, facilitating timely interventions.

The current system offers a cost-effective alternative to traditional diagnostic methods and is scalable for deployment in mobile or web applications. This makes it accessible for a broad audience, enhancing its potential for real-time health monitoring and personalized sleep recommendations.

Future enhancements include expanding the feature set to include additional factors like caffeine intake, screen time, and medication usage. Collecting larger and more diverse datasets will also improve the model's accuracy and generalization. Additionally, the system can be deployed as part of mobile apps, enabling continuous monitoring of sleep health and providing real-time feedback to users.

In summary, this project provides a solid foundation for future advancements in sleep health prediction, with the potential to transform how sleep disorders are diagnosed and managed through technology.
