|  |  |  |
| --- | --- | --- |
|  | G H Patel College of Engineering and Technology |  |

**COMPUTER ENGINEERING DEPARTMENT**

**Project Report on**

[**Sleep**](https://github.com/akshatkmistry/Parkinsons_Disease_Predictor-Voice_Measures) **disorder prediction**

**Submitted By**

**Name of Student-1: Daiwik Korat**

**Enrollment Number: 12202040501012**

**Name of Student-2: Aniket Mali**

**Enrollment Number: 12202040501007**

**Guided By: Dr. Priyang Bhatt**

**A.Y. 2024-25 EVEN TERM**

**Objective**  
The primary objective of this project is to develop an intelligent and accessible machine learning model capable of predicting the likelihood of sleep disorders based on an individual’s lifestyle and physiological parameters. Using a structured dataset, the system analyzes features such as age, sleep duration, screen time, stress levels, BMI category, blood pressure, and gender to assess the risk of sleep conditions like insomnia, sleep apnea, or general sleep deprivation. The project aims to serve as a preliminary diagnostic tool that leverages data analysis and classification techniques to assist in early detection of sleep-related health concerns. The entire model training, evaluation, and prediction pipeline has been implemented using Python within a Jupyter Notebook environment, emphasizing clarity, interpretability, and reproducibility.

**Dataset Used**

# For this project, we utilized the Sleep Health and Lifestyle Dataset ,available at [Kaggle](https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset)

**Key Features in Dataset**

1. **Age** – Age of the individual (in years).
2. **Gender** – Biological sex (Male/Female).
3. **Sleep Duration** – Average number of hours of sleep per night.
4. **Quality of Sleep** – Self-reported quality of sleep, typically on a scale (e.g., 1 to 10).
5. **Physical Activity Level** – Average physical activity level per day, usually in minutes.
6. **Stress Level** – Self-reported stress level on a scale from 1 to 10.
7. **BMI Category** – Category of Body Mass Index: Underweight, Normal Weight, Overweight, or Obese.
8. **Heart Rate** – Resting heart rate in beats per minute.
9. **Daily Steps** – Average number of steps taken per day.
10. **Screen Time** – Average screen time per day (in hours).
11. **Systolic Blood Pressure** – Upper number in a blood pressure reading (mmHg).
12. **Diastolic Blood Pressure** – Lower number in a blood pressure reading (mmHg).
13. **Alcohol/Smoking Addiction** – Presence or absence of addiction (Yes/No).
14. **Sleep Disorder** (Target Variable) – Indicates if the individual has a sleep disorder (None, Insomnia, or Sleep Apnea).

## Target Variable:

* **Sleep Disorder**  
  This is the main variable the model aims to predict. It has three possible classes like CRSD,Restless Leg Syndrome, Sleep Apnea, Insomnia, Parasomnia etc.

## Model Chosen:

We selected **Support Vector Machine (SVM)** for this project due to its effectiveness in handling classification problems with high-dimensional data. SVM works by finding the optimal hyperplane that separates the classes with the maximum margin, making it well-suited for health prediction tasks where the decision boundary needs to be precise. It performed well on our dataset in terms of accuracy, precision, and recall, making it a reliable choice for predicting potential sleep disorders based on patient data.

**How SVM Works**

**Support Vector Machine (SVM)** is a powerful supervised machine learning algorithm used primarily for classification tasks. The core idea behind SVM is to find the **optimal hyperplane** that best separates the data points of different classes.

In a two-dimensional space (for simplicity), this hyperplane is just a line that divides the space into two halves, each belonging to one class. But unlike basic classifiers, SVM doesn't just find any dividing line — it finds the one that **maximizes the margin** between the two classes. This margin is the distance between the hyperplane and the nearest data points from each class, which are called **support vectors**.

Here’s how it works step-by-step:

1. **Separation**: SVM looks for a line (or hyperplane in higher dimensions) that best separates the classes.
2. **Maximizing Margin**: Among all possible separating lines, SVM selects the one with the **largest margin** (maximum distance to the nearest points of each class).
3. **Support Vectors**: These are the data points that are closest to the hyperplane and directly influence its position and orientation.
4. **Kernel Trick (for non-linear data)**: If the data is not linearly separable, SVM uses a technique called the **kernel trick** to transform the data into a higher-dimensional space where a separating hyperplane *can* be found.

**Why SVM?**

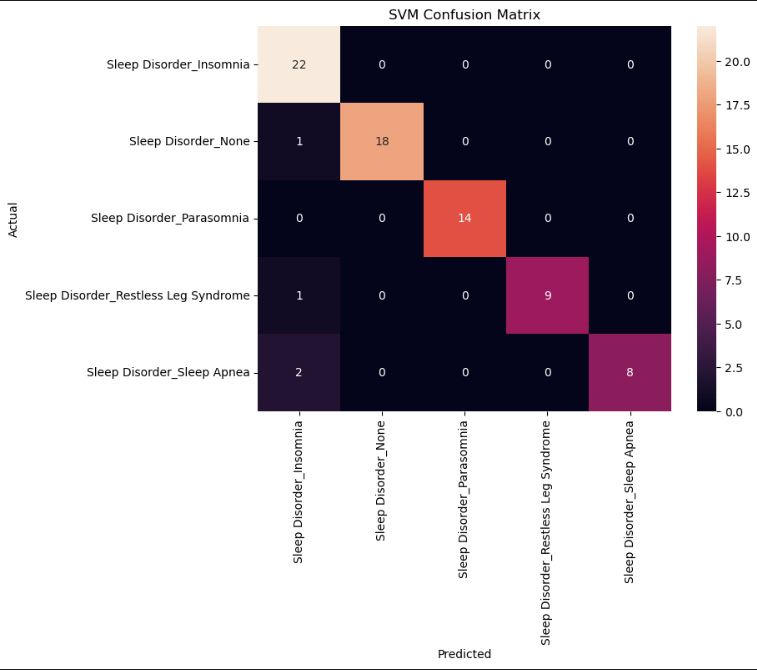
* We chose **Support Vector Machine (SVM)** for our Sleep Disorder Prediction project because of its high accuracy, efficiency, and ability to handle both linear and non-linear classification problems. SVM is particularly effective in high-dimensional spaces and works well when the number of features is greater than the number of samples — which is often the case in health-related datasets. It focuses on maximizing the margin between classes, which helps improve generalization and reduce overfitting. Additionally, SVM is robust to outliers and can perform well even with small datasets, making it an ideal choice for our project, which involved limited but meaningful medical and lifestyle data.

**Model Performance : -**

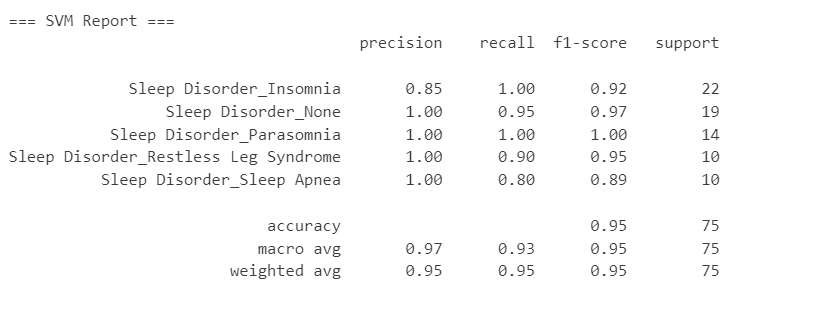
To assess the model's effectiveness, we employed the following metrics:

* Accuracy: Percentage of correct predictions
* Precision, Recall & F1-Score: Detailed evaluation of classification performance
* Confusion Matrix: Visualizes prediction distribution across classes

**Confusion Matrix:**



**Performance Metrics**

****

**Challenges & Learnings**

# Challenges Faced:

During the development of the Sleep Disorder Prediction model in Jupyter Notebook, several challenges were encountered. One of the primary difficulties was handling the mixed-type dataset, which included both numerical and categorical features. Encoding categorical variables like Gender and BMI Category in a way that preserved their meaning without causing bias was crucial. Another challenge was dealing with class imbalance in the target variable, which initially affected the model's accuracy. Selecting the right features also required careful analysis through correlation heatmaps and statistical techniques. Hyperparameter tuning of the SVM model was time-consuming, requiring multiple iterations to achieve optimal performance without overfitting. Ensuring the processed input matched the format expected by the model was also a subtle but important challenge.

# Key Learnings:

Working on this project enhanced our understanding of practical machine learning workflows within the Jupyter Notebook environment. We learned how to clean and preprocess healthcare data, handle missing values, and perform label encoding and feature scaling effectively. Experimenting with different models helped us appreciate why SVM was best suited for this problem — especially in terms of handling high-dimensional feature spaces. We also gained experience with evaluating model performance using various metrics such as accuracy, precision, recall, and F1-score. The project reinforced the importance of feature selection, proper data splitting, and understanding model behavior through confusion matrices and classification reports.

**Tools Used:** Python, Streamlit, Scikit-learn, Pandas, NumPy, **Matplotlib** & **Seaborn**

**Git Rep:** [Github](https://github.com/KoratDaiwik/ML-Project)

**References:**

1. A. Tsanas and M. A. Little, Machine Learning for Healthcare Technologies, London, U.K.: The Institution of Engineering and Technology (IET), 2016.
2. K. H. Yu and I. S. Kohane, Artificial Intelligence in Healthcare, Cambridge, MA, USA: MIT Press, 2021.