Sleep Disorder Prediction using Random Forest

Objective

The primary objective of this project is to develop a machine learning-based classification model capable of accurately predicting sleep disorders—namely Healthy, Sleep Apnea, and Insomnia—using health and lifestyle-related features.

This project uses the Random Forest Classifier, a robust ensemble model, and explores hyperparameter tuning using GridSearchCV to optimize performance. The aim is to compare the default model with the tuned version and determine which configuration performs best on the dataset. Ultimately, the goal is to deliver a model that not only performs well in terms of accuracy and classification metrics but is also reliable for real-world application in health analytics and early detection of sleep disorders.

Dataset Used

The dataset contains 15,000 rows and 13 columns representing data about individuals and their sleep-related health attributes.

Each row corresponds to a unique person, identified by a Person ID.

The target column is Sleep Disorder, which includes 3 categories:

- Healthy
- Sleep Apnea
- Insomnia

The data includes a mix of numerical and categorical features that can influence sleep disorders.

No missing values were found in the dataset — it is clean and ready for modeling.
The dataset includes a variety of useful columns:

- Gender: Male or Female
- Age: Age in years
- Occupation: Job title of the individual
- Sleep Duration: Number of hours the person typically sleeps
- Quality of Sleep: Rated on a scale of 1 (poor) to 10 (excellent)
- Physical Activity Level: Indicates how physically active the person is daily

- Stress Level: Rated on a scale of 1 (low) to 10 (high)
- BMI Category: Body Mass Index classification (e.g., Normal, Overweight)
- Blood Pressure: Given as a combined string like "120/80"
- Heart Rate: Beats per minute
- Daily Steps: Average number of steps walked in a day

The dataset provides a diverse representation of working professionals with various age groups and health statuses.

This real-world health data helps in identifying correlations between lifestyle choices and sleep disorders.

Model

Here the two models are used SVM and RandomForestClassifier.

- The performance of both the models are compared and the RandomForestClassifier has more performance than SVM.
- After that I have tried to increase the performance of the Random Forest Classifier using hyper parameter tuning.
- For hyper parameter tuning I have used GridSearchCV.This helped in finding the most suitable combination of parameters for the dataset.
- Additionally, to handle any potential class imbalance in the target labels, the parameter class_weight=balanced is used for balancing the dataset.
- After that the accuracy score of both default RandomForest and Tuned RandomForest is compared and the accuracy of the Tuned RandomForest comes slightly more so I have decided to use the Tuned RandomForest model for my project.

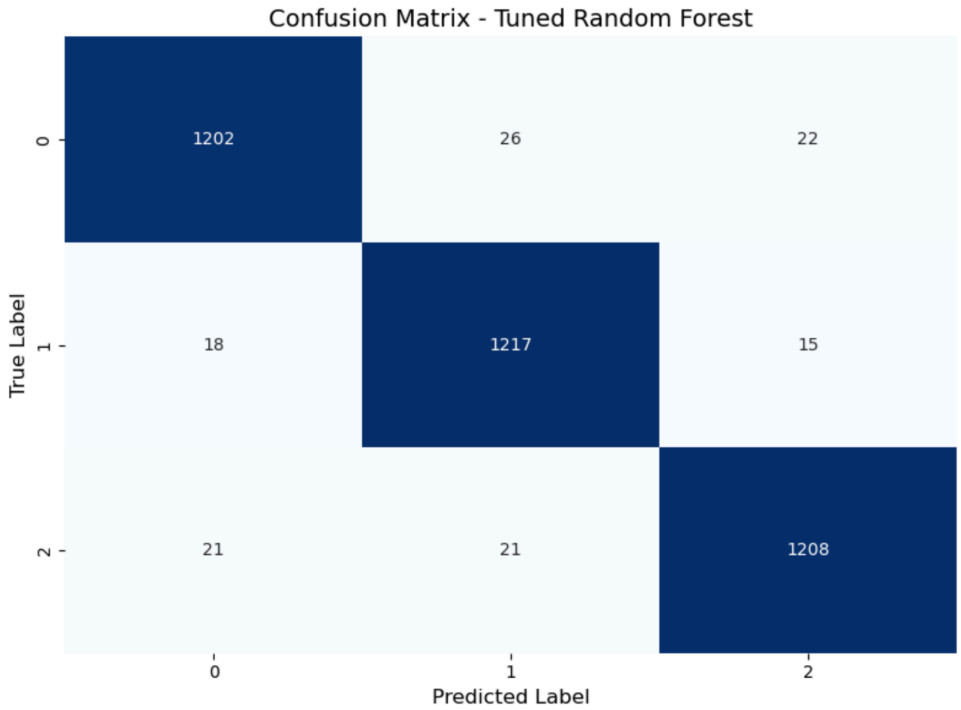
Performance Metrics

The model was trained using 75% of the data and tested on the remaining 25%. The following metrics were used to evaluate performance:

The detail are given below for Tuned RandomForest

- **Accuracy**
  - Training : 98.14%
  - Testing : 96.72%
- **F1-Score**:
  - Macro F1 score: 97%
  - Weighted F1 score: 97%

- **Confusion Matrix**: Minimal misclassifications, showing excellent generalization



The model performs exceptionally well across all classes, it is due to the well-distributed dataset and the power of ensemble learning.

Challenges

While working on this project, several challenges were encountered:

1. **Distinguishing Between Similar Classes:** It was difficult to correctly classify between *Healthy* and *Sleep Apnea* samples, even when their values appeared logically distinct.
2. **Model Bias Towards Dominant Features:** Initially, the model was overly reliant on specific features like *Stress Level* and *Quality of Sleep*, causing it to generalize less effectively.
3. **Hyperparameter Tuning:** After performing GridSearchCV for Random Forest tuning, the accuracy did not improve drastically and sometimes even slightly decreased.
4. **Finding the Right Feature Set:** Features like *Occupation* and *Diastolic* showed minimal impact on predictions.

Learnings

Through this project, I have gained a practical understanding of applying machine learning to real-world agricultural problems. Some key takeaways include:

1. Improved skills in **data preprocessing**, **exploratory data analysis**, and **feature selection**.

2. Learned how to build and deploy models using **Scikit-learn** and serve them using **Streamlit**.
3. Understood how **Random Forest** works internally and its advantages over other models in classification problems.
4. Understand how to perform hyperparameter tuning and what is the importance of hyperparameter tuning.