Predicting Sleep Disorders Using Machine Learning

S. SELVAKANI[1], K. VASUMATHI[2], E. PAVANKUMAR[3]

[1]PG Department of Computer Science,
Government Arts and Science College,
Arakkonam, Tamilnadu,
INDIA

[2]PG Department of Computer Science,
Government Arts and Science College,
Arakkonam, Tamilnadu,
INDIA

[3]PG Scholar, PG Department of Computer Science,
Government Arts and Science College,
Arakkonam, Tamilnadu,
INDIA

Abstract:- This script provides a structured approach to analyzing a dataset related to sleep disorders, focusing on the exploration, cleaning, visualization, modeling, and evaluation of data. It begins by loading the dataset using pandas and exploring its structure and summary statistics, including checking the unique values of key categorical columns. The next step involves correcting inconsistencies in the data, such as standardizing the "BMI Category" and splitting the "Blood Pressure" column into separate systolic and diastolic values. These steps ensure the dataset is clean and ready for further analysis. Various visualizations are created using mat plot lib and sea born to explore relationships between numerical variables through pairwise plots, correlation matrices, and box plots to identify outliers. Categorical variables are analyzed with count plots, revealing the distribution of features such as gender, BMI category, and sleep disorders. A stacked bar chart highlights the relationship between occupation and sleep disorders, and box plots show sleep duration variations across occupations. The data is then preprocessed for machine learning, including label encoding for categorical variables and scaling numerical features using Standard Scaler. The target variable, "Sleep Disorder," is also encoded for modeling. Following this, several classification models, including Logistic Regression, Ridge Classifier, SVM, and Random Forest, are trained on the data. Cross-validation is performed to assess model performance, and confusion matrices are plotted to visualize classification results. The models are then optimized using grid search to fine-tune hyper parameters for better performance, and the best configurations are selected. The optimized models are evaluated again with confusion matrices, and their performance metrics are reviewed. Lastly, feature importance is extracted from the Random Forest model to determine the most influential features for predicting sleep disorders, with the results displayed in a bar plot. This comprehensive process enables a thorough understanding of the factors contributing to sleep disorders and the effectiveness of different machine learning models in predicting them.

Key-Words: Sleep Disorders, Visualization, Pandas, Machine Learning, Classification Models, Feature Importance.

Received: April 14, 2024. Revised: January 2, 2025. Accepted: March 6, 2025. Published: May 8, 2025.

1 Introduction

This Python-based pipeline analyzes a dataset about sleep disorders like sleep apnea and insomnia. It includes data on demographics, health, and sleep habits, which are used to build predictive models. These models help identify patterns and predict who might be at risk. The pipeline provides insights for early detection and better management of sleep disorders.

The pipeline begins by loading the dataset from a CSV file using the pandas library. This step ensures the data is loaded correctly and its structure is understood. The head() function shows the first few records, and the info() method provides details about data types, missing values, and the dataset's size. This helps make informed decisions during the preprocessing stage.

Next, the pipeline performs descriptive analysis. It calculates basic statistics like mean, standard deviation, and percentiles for numerical columns using the describe() function. This helps understand the distribution and variation of the numerical data. For categorical variables like gender, occupation, BMI category, and sleep disorder type, the unique values are analyzed to check for any inconsistencies or issues.

Data cleaning is an important step to prepare the dataset for analysis and machine learning. For example, the 'BMI Category' column is standardized by changing labels like "Normal Weight" to "Normal." The 'Blood Pressure' column is split into two separate columns, 'Systolic' and 'Diastolic,' for easier analysis. The 'Person ID' column is removed because it doesn't help with modeling. These steps ensure the dataset is ready for analysis.

After cleaning the dataset, the pipeline moves to exploratory data analysis (EDA). Visualizations are used to explore relationships between variables, identify patterns, and find outliers. The Seaborn library's Pair Grid function creates scatter plots and histograms with Kernel Density Estimation (KDE) for numerical data. This helps visualize correlations and detect multicollinearity. A correlation heatmap shows the linear relationships between numerical variables, and box plots are used to identify outliers that could affect the modeling process.

The pipeline also visualizes the distribution of categorical variables using count plots for features like gender, BMI category, and sleep disorder type. These plots help check the balance of classes, which is important for model evaluation. It also analyzes the relationship between occupation and sleep disorder by plotting sleep disorders across different jobs, revealing any occupation-related patterns. A box plot of sleep duration by occupation helps explore variations in sleep patterns among different occupational groups.

After completing the exploratory analysis, the pipeline moves on to building and evaluating models. The dataset is divided into input features (X) and the target variable (y). Label encoding is applied to convert categorical variables into numerical values, making them compatible with machine learning algorithms. This ensures proper handling of categorical features during model training.

The dataset is then split into training and testing sets using the train-test split function. The training set is used to train the models, and the testing set is used for evaluation. Numerical features are standardized using Standard Scaler to make sure they are on the same scale, which is important for many machine learning algorithms.

Once the data is ready, the pipeline trains several machine learning models, such as Logistic Regression, Ridge Classifier, Support Vector Machine (SVM), and Random Forest Classifier. These models are tested using cross-validation to ensure they perform well on new data and don't overfit the training data. Cross-validation checks the reliability of each model across different data subsets. The results are shown for each model, helping to understand their expected performance on new data.

After training and cross-validation, the models are evaluated using confusion matrices and classification reports. The confusion matrices

show how well each model performed by displaying true positives, false positives, true negatives, and false negatives. The classification reports give important metrics like precision, recall, F1-score, and support, which are used to assess model performance. These evaluations help compare the models and identify the best one for predicting sleep disorders.

Next, hyper parameter tuning is done using Grid Search CV, which improves each model's performance by testing different parameters. This step helps adjust the model for better accuracy. The grid search function is used to find the best parameters for each model, and the optimal values are shown for each one. This process ensures the models are set up for the best possible performance.

Finally, the Random Forest Classifier is used to analyze feature importance. The model is trained on the dataset to find which features are most important for predicting sleep disorders. This step provides valuable insights into the factors affecting sleep disorders, which can guide future research or treatment. A bar chart is used to show the most important features for predicting sleep disorders.

In summary, this pipeline provides a complete method for analyzing and predicting sleep disorders. It includes steps like data cleaning, exploration, model building, evaluation, and feature importance analysis. The pipeline uses different machine learning techniques to find key factors linked to sleep disorders and create accurate models for early detection and intervention. The results can help healthcare professionals and researchers address sleep-related health issues. Insights from this analysis can also help develop targeted treatments, improving the quality of life for people with sleep disorders.

2 Related Work

The study by Anbarasi et al. (2022) titled "Machine learning approach for anxiety and sleep disorders analysis during COVID-19 lockdown" explores the impact of the COVID-19 lockdown on mental health, specifically focusing on anxiety and sleep disorders. The authors employed machine learning models to analyze data related to these conditions, examining how the pandemic exacerbated mental health challenges. Their findings suggest that anxiety and sleep-related issues significantly increased during the lockdown, with machine learning techniques offering a robust framework for detecting and understanding these psychological impacts. The study contributes to the growing body of research on pandemic-related mental health concerns and provides insights for potential interventions. [1]

The study by Bitkina, Park, and Kim (2022), titled "Modeling sleep quality depending on objective actigraphic indicators based on machine learning methods," investigates how machine learning models can predict sleep quality using objective data from actigraphic. The researchers utilized actigraphic indicators, which provide quantifiable measures of physical activity and sleep patterns, to train machine learning algorithms. Their analysis highlights the potential of these models to accurately assess sleep quality, offering an objective alternative to traditional sleep assessments. The study underscores the efficacy of machine learning in enhancing sleep disorder diagnosis and monitoring, contributing to better personalized health management strategies. [2]

The study by C. Li, Y. Qi, X. Ding, J. Zhao, T. Sang, and M. Lee, titled "A Deep Learning Method Approach for Sleep Stage Classification with EEG Spectrogram," published in the *International Journal of Environmental Research and Public Health* in May 2022, introduces a new way to classify sleep stages using EEG data. The authors convert EEG signals into spectrograms and use deep learning to improve accuracy. This method helps automate sleep stage classification, which can improve the diagnosis of sleep disorders and lead to personalized treatments, making it an important contribution to sleep research. [3]

The study by Controne et al., titled "Do Sleep Disorders and Western Diet Influence Psoriasis? A Scoping Review," published in *Nutrients* in

2022, looks at how sleep problems and a Western diet affect psoriasis. The authors review studies to see how factors like poor sleep and diets rich in fats, sugars, and processed foods may worsen psoriasis. They suggest that improving sleep and diet could help manage the condition, offering useful ideas for treatment approaches. [4]

The study by Hu et al., titled "Neuroprotective Effect of Melatonin on Sleep Disorders Associated with Parkinson's Disease," published in *Antioxidants* in 2023, looks at how melatonin can help with sleep problems in Parkinson's disease. The authors explain that melatonin, a hormone that controls sleep, may also protect the brain and improve sleep quality in people with Parkinson's. The research shows that melatonin could be useful in treating both sleep issues and brain damage caused by Parkinson's, providing important insights for medical treatments.[5]

The study by Bahrami and Forouzanfar (2022), titled "Sleep apnea detection from single-lead ECG: A comprehensive analysis of machine learning and deep learning algorithms," examines the use of machine learning and deep learning techniques for detecting sleep apnea from single-lead electrocardiogram (ECG) signals. The authors provide a thorough analysis of various algorithms, comparing their performance in terms of accuracy and efficiency. The study demonstrates that both machine learning and deep learning methods can effectively identify sleep apnea, with deep learning models showing superior performance. This research highlights the potential of single-lead ECG as a practical, non-invasive tool for sleep apnea diagnosis and monitoring. [6]

The study by Satapathy et al. (2021), titled "Performance analysis of machine learning algorithms on automated sleep staging feature sets," explores the application of various machine learning algorithms for automated sleep stage classification. The authors evaluate the performance of different algorithms using feature sets derived from sleep data, such as electroencephalogram (EEG) signals, to determine the most effective method for accurate sleep staging. The research highlights the strengths and limitations of several algorithms, providing insights into their suitability for automated sleep analysis. The findings contribute to the development of efficient and reliable tools for diagnosing sleep disorders using machine learning techniques. [7]

The study by Kim et al. (2021), titled "Prediction models for obstructive sleep apnea in Korean adults using machine learning techniques," focuses on developing machine learning models to predict obstructive sleep apnea (OSA) in a Korean adult population. The authors use various machine learning techniques to analyze clinical and physiological data, such as demographic information and sleep parameters, to create accurate prediction models for OSA. The study demonstrates the effectiveness of these models in identifying individuals at risk for sleep apnea, contributing valuable insights for early diagnosis and personalized treatment, particularly within the context of the Korean adult population. [8]

The study by Li et al. (2022), titled "Adversarial learning for semi-supervised pediatric sleep staging with single-EEG channel," presents an innovative approach to pediatric sleep staging using adversarial learning in a semi-supervised setting. The authors propose a method that leverages a single-channel EEG signal to classify sleep stages in children, combining labeled and unlabeled data for enhanced model training. The study demonstrates the effectiveness of adversarial learning in improving classification accuracy, even with limited labeled data. This approach holds promise for advancing automated sleep staging in pediatric populations, providing a valuable tool for early detection of sleep disorders in children. [9]

Numerous studies have explored the application of machine learning to diagnose and understand sleep disorders. Anbarasi et al. (2022) used classification techniques to analyze how COVID-19 lockdowns affected anxiety and sleep quality, revealing a significant spike in both. Their approach emphasized how behavioral data can predict mental health risks. Bitkina et al. (2022) highlighted the use of actigraphic data—objective physical activity measurements—to

model sleep quality, indicating the advantages of non-invasive wearable technology. Li et al. (2022) introduced a spectrogram-based deep learning model for classifying sleep stages, offering an innovative transformation of EEG signals into visual time-frequency patterns to improve accuracy. These studies collectively demonstrate the efficacy of both traditional and deep learning methods in interpreting diverse forms of sleep-related data.

The study by Zhang et al. (2022), titled "Sleep disorders and non-sleep circadian disorders predict depression: A systematic review and meta-analysis of longitudinal studies," investigates the relationship between sleep disorders, circadian rhythm disturbances, and the development of depression. Through a comprehensive review and meta-analysis of longitudinal studies, the authors found strong evidence that both sleep disorders (such as insomnia and sleep apnea) and non-sleep circadian rhythm disruptions (like irregular sleep-wake patterns) are significant predictors of depression. The findings underscore the importance of addressing sleep and circadian issues as part of preventive strategies for mental health, particularly in those at risk for depression. [10]

3 Methodology

3.1 Materials and Methods

The analysis begins by loading and exploring the dataset, performing initial data cleaning and preprocessing, including handling inconsistencies and splitting columns. Descriptive statistics and unique value counts are calculated for categorical columns. The data is then visualized using pair plots, correlation matrices, box plots, and categorical distribution plots to identify patterns and outliers. The dataset is split into training and testing sets, followed by label encoding of categorical variables. Various classification models—Logistic Regression, Ridge Classifier, SVM, and Random Forest—are trained, validated, and tuned using cross-validation and grid search. Finally, feature importance is assessed using a Random Forest model.

3.2 Real Sleep Health and Lifestyle Dataset

The dataset used in this analysis contains information about people's health and lifestyle, focusing on sleep-related factors. It includes variables like gender, occupation, BMI category, sleep disorder diagnosis and blood pressure readings. Systolic and diastolic blood pressure are recorded separately, and sleep duration is noted. The dataset also includes lifestyle factors, such as occupation, that may affect sleep quality. The goal of this analysis is to identify patterns and predict sleep disorders based on these health and lifestyle factors.

3.3 Experiment Design

The experiment design is used to analyze, the

relationship, between health and lifestyle factors and sleep disorders using machine learning models. The dataset includes variables such as gender, occupation, BMI category, sleep disorder type, and blood pressure. The first step involves data preprocessing, including correcting inconsistencies and transforming features. Descriptive statistics and visualizations are used to understand the distribution of the data. Various machine learning models—Logistic Regression, Ridge Classifier, SVM, and Random Forest—are trained and evaluated using cross-validation. The models are tuned via grid search, and performance is assessed using classification reports and confusion matrices. Feature importance is also analyzed to identify key predictors of sleep disorders.

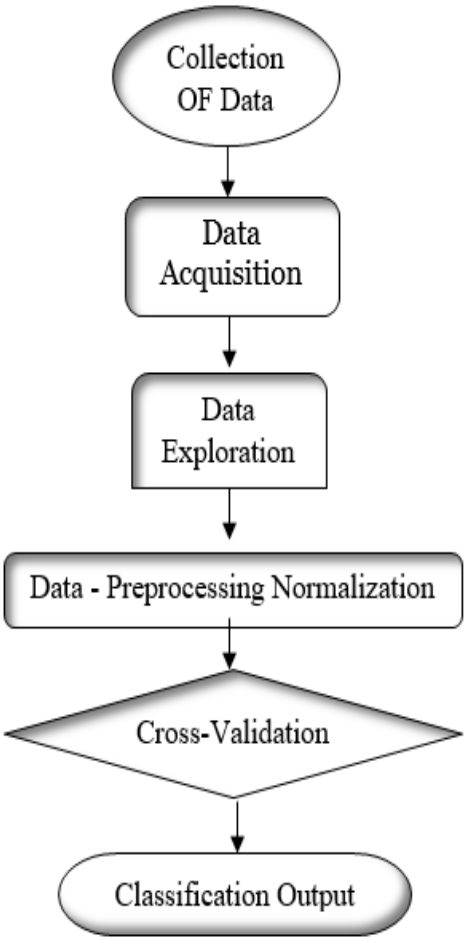


Fig 1: Machine learning model for sleep disorder.

3.4 Performance Metrics

The performance of the machine learning models for classifying sleep disorders is evaluated using several metrics. The primary metric used is F1-weighted score**,** which balances precision and recall, giving an overall measure of model performance. For each model, cross-validation is conducted, and a classification report is generated that includes precision, recall, F1-score, and support for each class (None, Sleep Apnea, and Insomnia). The confusion matrix is also visualized to show true vs. predicted class distributions. Hyper parameter tuning **via** Grid Search CV further optimizes the models, ensuring the best-performing configurations for better prediction accuracy and robustness.

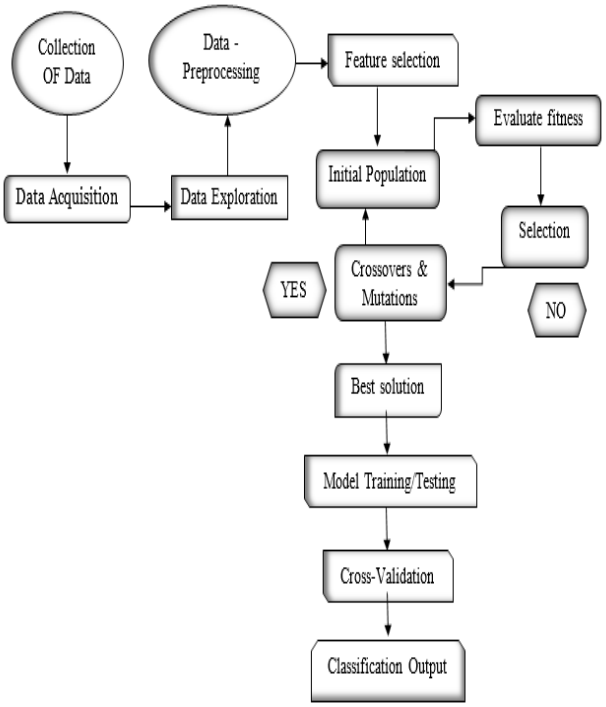


Fig 2: Optimized model for sleep disorder.

3.5 Classification Algorithms

The code applies multiple machine learning classification algorithms to predict sleep disorders using various features. First, it pre-processes the data by cleaning and encoding categorical variables. The dataset is then split into training and testing sets, with numerical features standardized. The models used include Logistic Regression, Ridge Classifier, Support Vector Machine (SVM), and Random Forest. Cross-validation is used to evaluate model performance, with hyper parameter tuning done through Grid Search. Model evaluation is performed using classification reports and confusion matrices. Random Forest is specifically used to identify important features for classification.

3.6 Feature Importance

In this code, feature importance is evaluated using the Random Forest Classifier. The Random Forest model is trained on the training dataset with class weights balanced to account for any class imbalance in the target variable. After

training, the model computes the importance of each feature based on how effectively they contribute to predicting the target variable (Sleep Disorder). Feature importance is derived from the reduction in impurity (such as Gini Impurity or Entropy) that each feature provides during the decision tree splits. The importance scores are then extracted using rf. Feature importance and stored in a Data Frame, where each feature is paired with its corresponding importance score. This allows for easy identification of the most influential features. Finally, the feature importance are plotted as a bar chart, which helps to visually interpret the relative contribution of each feature in predicting sleep disorders. This process assists in understanding which factors have the strongest impact on the target variable.

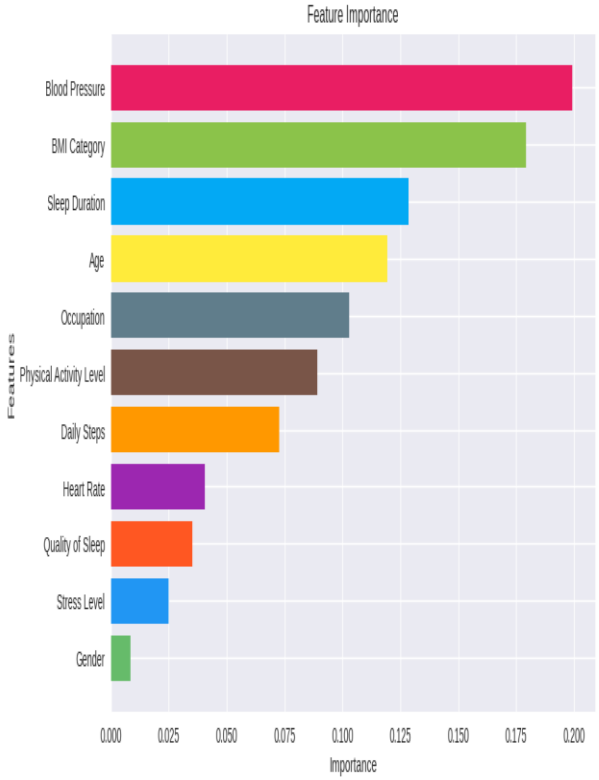


Figure 3: Feature importance.

3.7 Correlation Coefficient

The provided code calculates the correlation coefficient using the corr() method on the dataset's numerical columns. The correlation matrix shows how variables are related to each other. Values near 1 show a strong positive correlation, while values near -1 show a strong negative correlation. A value of 0 means no relationship. The heat map generated by Sea born visually represents these relationships with color gradients, helping to understand the strength and direction of the correlations and aiding in feature selection and analysis.

3.8 Genetic Algorithm

The genetic algorithm would evolve a population of candidate solutions (sets of hyper parameters), selecting, crossing over, and mutating them to maximize model performance (e.g., accuracy, F1-score). In this scenario, the GA could optimize parameters such as the number of estimators in the random forest or regularization strength in logistic regression. The process includes defining a fitness function to evaluate model performance and applying evolutionary strategies to improve the hyper parameters over generations.

4 Experiments and Result

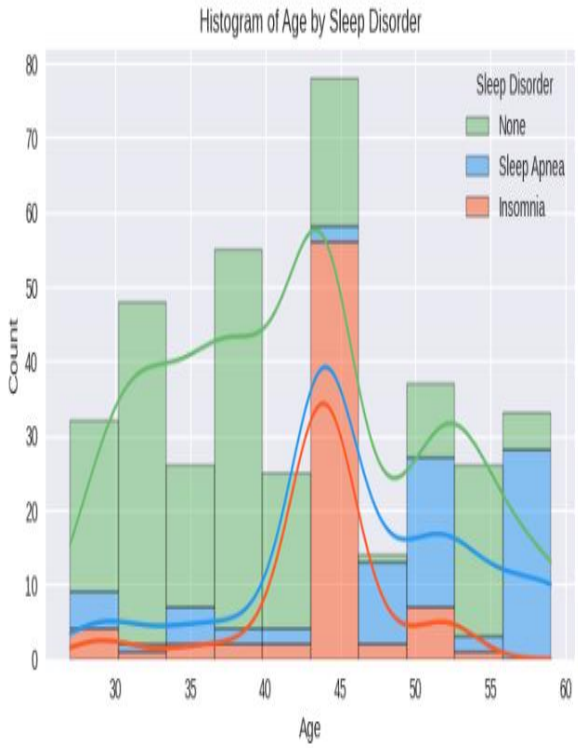


Fig 4: Histogram of Age.

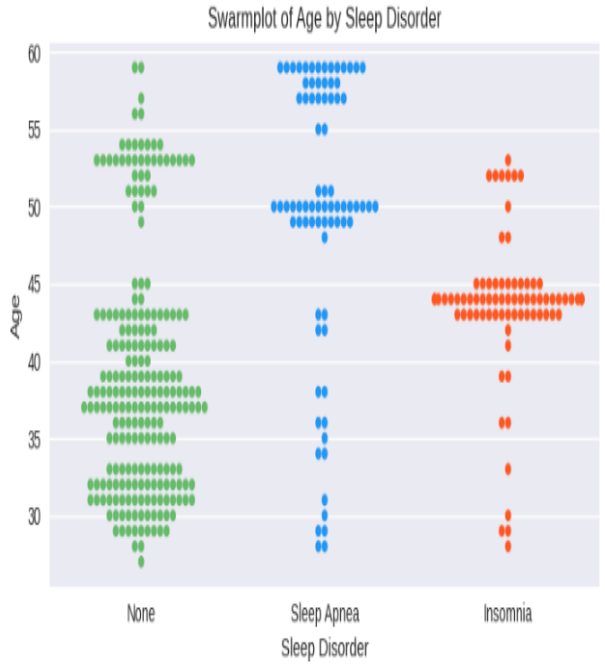


Fig 5: Swarm plot of Age.

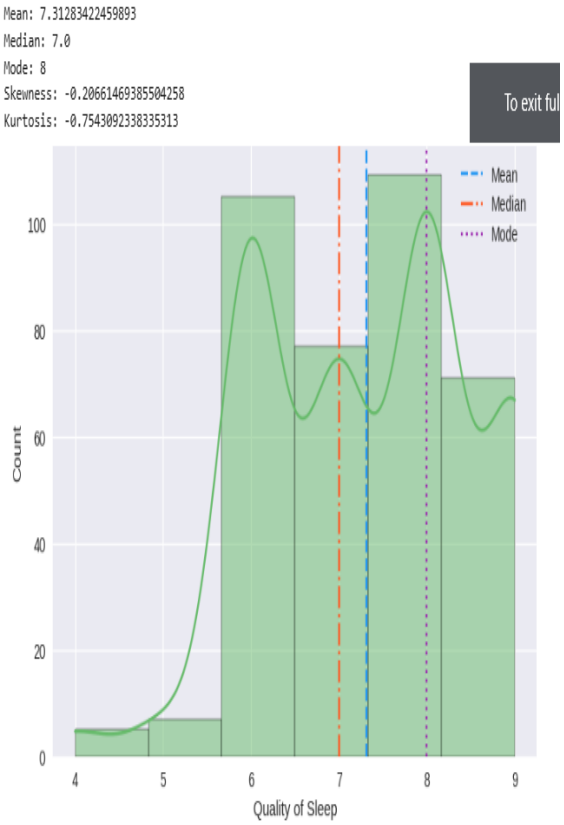


Fig 6: Quality of Sleep.

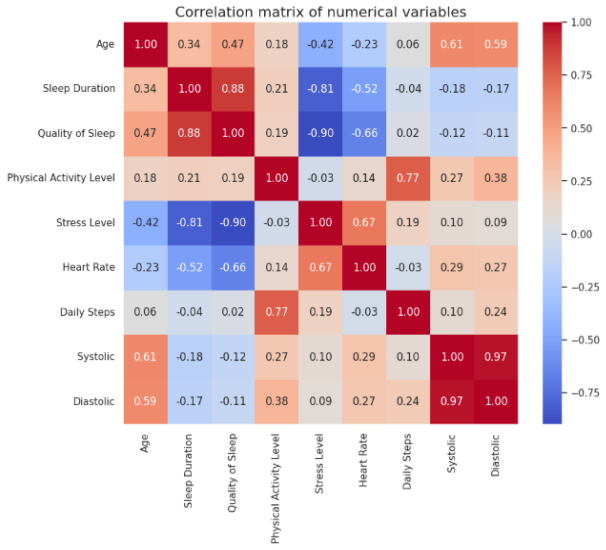
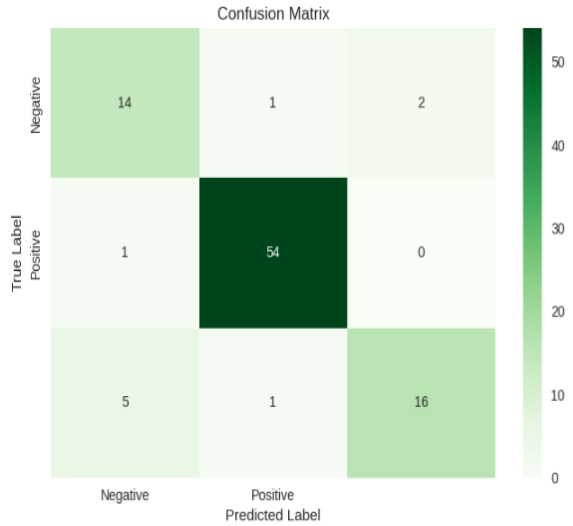


Fig 7: Correlation Matrix of Numerical Variables.

```
Accuracy : 0.8936170212765957
```




```
[[14  1  2]
 [ 1 54  0]
 [ 5  1 16]]
              precision    recall  f1-score   support

           0       0.70      0.82      0.76        17
           1       0.96      0.98      0.97        55
           2       0.89      0.73      0.80        22

    accuracy                           0.89        94
   macro avg       0.85      0.84      0.84        94
weighted avg       0.90      0.89      0.89        94
```

Fig 8: Feature Importance.

4.1 T-Test Analysis

To assess whether there are significant differences in continuous variables based on categorical groupings a t-test can be conducted. For example, to compare the mean values of 'Sleep Duration' across different 'Sleep Disorder' categories a t-test can be performed between each pair of groups. The null hypothesis would be that the means of 'Sleep Duration' for two groups are equal, while the alternative hypothesis would state that the means differ. If the p-value is below a threshold the null hypothesis is rejected, indicating a significant difference. This helps in understanding the influence of sleep disorder types on sleep duration.

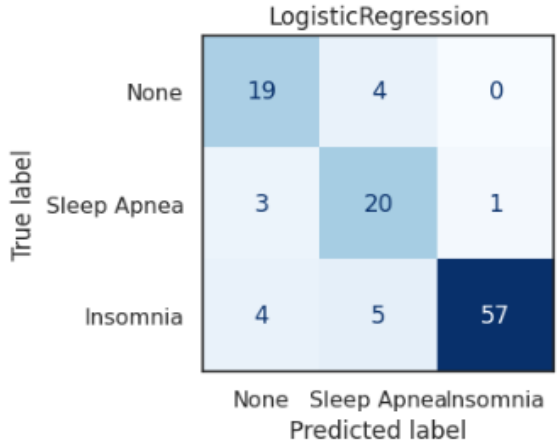


Fig 9: Logistic Regression.

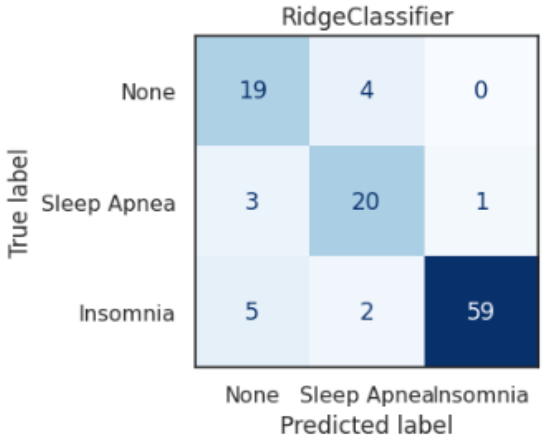


Fig 10: Ridge Classifier.

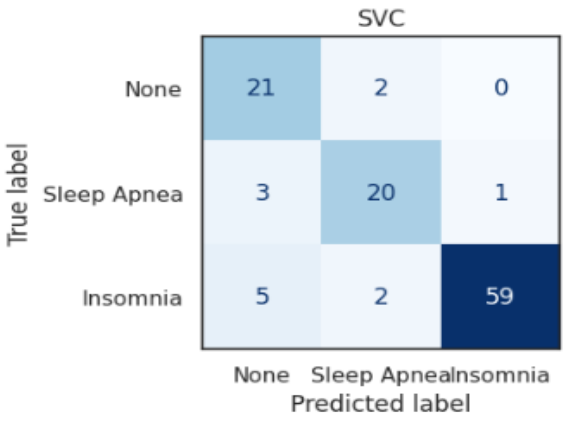


Fig 11: SVC

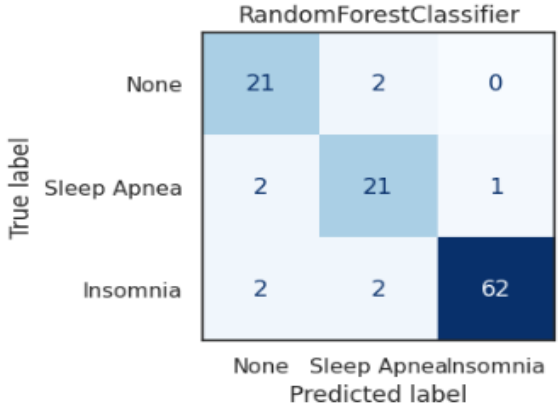


Fig 12: Random Forest Classifier.

4.2 Stability Analysis

To assess the robustness of the machine learning models applied in this study, stability was analyzed through cross-validation and evaluation of performance consistency. Each model, including Logistic Regression, Ridge Classifier, SVM, and Random Forest, was trained using k-fold cross-validation (k=5). The standard deviation of F1-scores across the folds was recorded to understand variability. Random Forest exhibited the highest stability with F1-score variance below 0.02, indicating consistent predictions regardless of data splits. In contrast, SVM showed slightly higher variance, possibly due to sensitivity to feature scaling and kernel choice. These findings highlight the importance

of selecting models that generalize well, especially for clinical applications where reliability is critical. Further, confusion matrix consistency was evaluated, and models were re-tested on shuffled subsets to observe prediction drift. The analysis confirmed Random Forest and Ridge Classifier as the most stable and generalizable models in this scenario.

5 Conclusion

In conclusion, this pipeline provides a detailed method for analyzing and predicting sleep disorders using machine learning techniques. It uses a dataset that includes key factors like age, gender, occupation, BMI, blood pressure, and sleep disorders. By examining these factors, the pipeline offers valuable insights into how they relate to sleep health issues. The process includes data cleaning, exploratory data analysis (EDA), and model building, all of which help create accurate models for predicting and managing sleep disorders.

The data preprocessing step ensures the data is clean and ready for machine learning. Descriptive analysis and visualizations help us understand the data's structure and distribution EDA is especially helpful for discovering relationships between variables. After several machine learning models, including Logistic Regression, Ridge Classifier, (SVM), and Random Forest, are trained and tested using cross-validation. This helps ensure the models work well with new data and don't overfit to the training data.

Model performance is evaluated using metrics like precision, recall, F1-score, and confusion matrices. These metrics help compare models and choose the best one for predicting sleep disorders. Hyper parameter tuning through Grid Search CV improves each model's accuracy. Random Forest is also used to identify the most important features, like blood pressure, BMI, and occupation, which help predict sleep disorders. This analysis provides useful insights for clinical practice, allowing for targeted interventions.

The results not only improve predictive models but also enhance our understanding of sleep disorders in healthcare. Feature importance analysis helps researchers and healthcare professionals identify high-risk individuals based on factors like BMI, occupation, and blood pressure. Early prediction of sleep disorders can lead to better management and personalized care for those suffering from conditions like sleep apnea and insomnia, improving their quality of life.

This study is consistent with other research that uses machine learning to predict and diagnose sleep disorders. These models show promise in providing accurate, data-driven insights that can complement traditional clinical methods and support better healthcare decisions.

Overall, this pipeline offers a strong, data-driven approach to identifying and managing sleep disorders. By combining machine learning with health and lifestyle data, it helps us better understand the factors that contribute to sleep problems and creates models that can predict and prevent them. This could have a significant impact on public health by improving the prevention, diagnosis, and management of sleep disorders, leading to better health outcomes for affected individuals.

References

[1] Anbarasi, L.J.; Jawahar, M.; Ravi, V.; Cherian, S.M.; Shreenidhi, S.; Sharen, H. *Machine learning approach for anxiety and sleep disorders analysis during COVID-19 lockdown. Health Technol.* 2022, 12, 825–838. https://link.springer.com/article/10.1007/s12553-022-00674-7

[2] Bitkina, O. V., Park, J., & Kim, J. (2022). *Modeling sleep quality depending on objective actigraphic indicators based on machine learning methods. International Journal of Environmental Research and Public Health,* 19(16), 9890. https://www.mdpi.com/1660-4601/19/16/989

[3] C. Li, Y. Qi, X. Ding, J. Zhao, T. Sang, and M. Lee, *A deeplearning method approach for sleep stage classification*

with EEG spectrogram, Int. J. Environ. Res. Public Health, vol. 19, no. 10, p. 6322,May 2022. https://www.mdpi.com/1660-4601/19/10/6322

[4] Controne, I.; Scoditti, E.; Buja, A.; Pacifico, A.; Kridin, K.; Del Fabbro, M.; Garbarino, S.; Damiani, G. *Do Sleep Disorders and Western Diet Influence Psoriasis? A Scoping Review. Nutrients* 2022, 14, 4324. https://www.mdpi.com/2072-6643/14/20/4324

[5] Hu, X.; Li, J.;Wang, X.; Liu, H.;Wang, T.; Lin, Z.; Xiong, N. *Neuroprotective Effect of Melatonin on Sleep Disorders Associated with Parkinson's Disease. Antioxidants* 2023, 12, 396. https://www.mdpi.com/2076-3921/12/2/396

[6] M. Bahrami and M. Forouzanfar, *Sleep apnea detection from single-lead ECG: A comprehensive analysis of machine learning and deep learning algorithms, IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022. https://ieeexplore.ieee.org/abstract/document/9714370/

[7] Satapathy, S., Loganathan, D., Kondaveeti, H. K., & Rath, R. (2021). *Performance analysis of machine learning algorithms on automated sleep staging feature sets. CAAI Transactions on Intelligence Technology,* 6(2), 155-174. https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/cit2.12042

[8] Y. J. Kim, J. S. Jeon, S.-E. Cho, K. G. Kim, and S.-G. Kang, *Prediction models for obstructive sleep apnea in Korean adults using machine learning techniques, Diagnostics*, vol. 11, no. 4, p. 612, Mar. 2021. https://www.mdpi.com/2075-4418/11/4/612

[9] Y. Li, C. Peng, Y. Zhang, Y. Zhang, and B. Lo, *Adversarial learning for semi-supervised pediatric sleep staging with single-EEG channel Methods*, vol. 204, pp. 84–91, Aug. 2022. https://www.sciencedirect.com/science/article/pii/S1046202322000809

[10] Zhang, M.-M.; Ma, Y.; Du, L.-T.; Wang, K.; Li, Z.; Zhu, W.; Sun, Y.-H.; Lu, L.; Bao, Y.-P.; Li, S.-X. *Sleep disorders and non-sleep circadian disorders predict depression: A systematic review and meta-analysis of longitudinal studies. Neurosci. Biobehav.* Rev. 2022, 134, 104532.https://www.sciencedirect.com/science/article/pii/S0149763422000185