# 

**CHAPTER 1**

# Introduction to Organisation

BML Munjal University (BMU), located in Gurugram, Haryana, is a prestigious institution known for its commitment to academic excellence and innovative learning. Established in 2014, BMU aims to nurture ethical leaders who are skilled, knowledgeable, and capable of creating a sustainable future.

With a strong emphasis on interdisciplinary education and hands-on learning, the university has earned recognition for its programs in engineering, management, and research. BMU is equipped with state-of-the-art facilities, fostering a thriving environment for innovation and exploration.

The university's focus on cutting-edge research in Artificial Intelligence, Data Analytics, and Health Sciences provided a robust platform for the development of the **"Sleep Disorder Prediction"** project. The mentorship, technical support, and access to specialized datasets offered by BMU significantly contributed to the project's successful completion.

**CHAPTER 2**

# Introduction to Project

## Overview

Sleep disorders are a critical concern in healthcare, significantly impacting individuals' quality of

life. Identifying and diagnosing these disorders often requires extensive data collection and

analysis, which can be subjective and time-consuming. This project leverages machine learning

to predict sleep disorders by analyzing patterns and relationships within a dataset of lifestyle and

health factors.

The core idea of this project is to use supervised learning techniques to identify patterns in data

related to sleep health, such as sleep duration, stress levels, and BMI categories. By examining

these factors, the model predicts the likelihood of sleep disorders like Insomnia or Sleep Apnea.

This approach automates the analysis, reducing subjectivity and improving diagnostic efficiency.

The model development process includes data preprocessing, feature engineering, and training

classification algorithms. Python was chosen for its robust libraries and flexibility in data

manipulation and machine learning model implementation. The insights generated by this

project aim to streamline sleep disorder diagnosis and contribute to better resource allocation in

healthcare.

## Existing System

## The traditional system for diagnosing sleep disorders relies on clinical methods such as patient

## interviews, questionnaires, and sleep studies like polysomnography. These approaches, while

## effective, are time-consuming, costly, and limited in accessibility. Diagnosis depends heavily on

## subjective assessments by healthcare professionals, which may lead to variability in results.

## Moreover, the manual process of data interpretation and the high cost of in-lab studies make

## these methods impractical for large-scale use.

## User Requirement Analysis

**Functional Requirements**: The system must predict sleep disorders (None, Insomnia, Sleep

Apnea) based on user inputs such as age, BMI, sleep duration, and stress levels. Clear

visualizations and high prediction accuracy are essential.

**Non-Functional Requirements**: The system should be scalable, secure, and user-friendly,

offering real-time predictions. Compatibility with healthcare systems and portability across

platforms are also required.

**Stakeholder Expectations**: Healthcare professionals need a reliable diagnostic aid, researchers

require robust data analysis, and individuals seek affordable, accessible sleep disorder insights.

## Feasibility Study

**Technical Feasibility**:

The project utilizes Python and its robust libraries for machine learning, ensuring efficient data

processing and analysis. A comprehensive dataset with relevant features is available, and

classification algorithms like Random Forest and Decision Tree are well-suited for accurate

predictions.

**Operational Feasibility**:

The system addresses a real-world need, offering a faster and more affordable alternative to

traditional diagnostic methods. It is user-friendly, scalable, and reduces reliance on manual, time-

intensive processes.

**Economic Feasibility**:

Development and deployment costs are minimized through open-source tools and scalable cloud-

based solutions, making the system cost-effective and accessible.

**Legal and Ethical Feasibility**:

The system adheres to data privacy regulations and incorporates measures to ensure unbiased

predictions.

**Schedule Feasibility**:

The project is achievable within the given timeline using agile methodologies for iterative

development and testing.

This feasibility study confirms the project is practical, sustainable, and impactful.

**CHAPTER 3**

# Literature Review

The analysis and prediction of sleep disorders have been critical areas of research in healthcare. Researchers have historically relied on manual techniques to identify trends and relationships in sleep-related data, which is both labor-intensive and subject to human bias. The evolution of technology, particularly machine learning, has paved the way for automated and efficient methods of analysis.

**Identification of Key Determinants**

Early studies identified factors like BMI, stress levels, and physical activity as significant determinants of sleep health. For example, Evangelopoulos et al. (2012) used latent semantic analysis to explore the relationships between various health parameters and their impact on sleep disorders. Their work highlighted the importance of integrating diverse data sources to uncover meaningful patterns. Similarly, Canini et al. (2009) emphasized the role of unsupervised methods in detecting latent variables in health datasets.

**Machine Learning in Sleep Disorder Prediction**

Machine learning techniques have shown great promise in sleep disorder diagnostics. Hurtado et al. (2016) demonstrated the use of clustering and classification techniques for analyzing sleep data, achieving significant improvements in diagnostic efficiency. Q. Mei and C. Zhai (2005) introduced temporal data mining methods, which provided new insights into time-dependent patterns in sleep studies. Their research laid the groundwork for integrating time-series analysis into predictive models.

Saini et al. (2013) advanced this field by applying ensemble methods like Random Forest for sleep disorder classification. Their results indicated that ensemble methods outperform single classifiers by reducing overfitting and improving prediction accuracy.

**3.1 Comparative Analysis of Techniques**

A comparative study of various machine learning algorithms for sleep disorder prediction showed that Decision Tree and Random Forest models consistently performed well. Random Forest was particularly noted for its robustness and ability to handle noisy data effectively. Table 1 below summarizes the performance metrics of these models from previous studies:

| **Research Study** | **Algorithm** | **Accuracy** | **Key Findings** |
| --- | --- | --- | --- |
| Evangelopoulos et al. (2012) | Latent Semantic Analysis | 75% | Highlighted latent factors affecting sleep. |
| Saini et al. (2013) | Random Forest | 90% | Reduced overfitting, robust to noisy data. |
| Hurtado et al. (2016) | Clustering & Classification | 85% | Efficient classification of sleep disorders. |

**3.2 Objectives of the Project**

The reviewed literature reveals several gaps and opportunities:

1. Traditional methods are time-consuming and lack scalability, necessitating automated solutions.
2. While machine learning has been effective, many studies overlook the integration of comprehensive datasets that include demographic, lifestyle, and physiological factors.

**Building on these insights, the objectives of this project are:**

* To develop a machine learning-based model capable of predicting sleep disorders (None, Insomnia, Sleep Apnea) with high accuracy.
* To analyze relationships between factors like BMI, stress levels, and occupation in determining sleep health.
* To provide a scalable and cost-effective solution that complements traditional diagnostic methods.

**CHAPTER 4**

# Exploratory Data Analysis

EDA was conducted to understand the dataset and identify patterns relevant to

sleep disorder prediction. The dataset comprises 400 entries and 13 features,

including demographic, lifestyle, and health-related variables. The target variable,

"Sleep Disorder," has three categories: None, Insomnia, and Sleep Apnea.

**1. Dataset Overview**

* Key Features: Age, gender, occupation, sleep duration, quality of sleep, stress
* level, BMI, and daily steps.
* Statistics:
  + Most individuals are aged 30–45 years.
  + Sleep duration averages 5–7 hours.
  + Normal BMI is the most common, followed by overweight and obese.
  + A majority reported no sleep disorders, with smaller proportions having
  + Insomnia or Sleep Apnea.

**2. Data Cleaning and Transformation**

* Missing values in "Sleep Disorder" were replaced with "None."
* Blood pressure was split into systolic and diastolic components.
* BMI categories were standardized for consistency.

**3. Key Findings from Analysis**

* **Gender and Sleep Disorders**: Females exhibited more Sleep Apnea cases, while Insomnia was higher among males.
* **Occupation Impact**: Nurses and sales representatives had higher instances of

disorders due to stress and irregular schedules.

* **BMI and Disorders:** Overweight individuals were more prone to sleep

disorders, particularly Sleep Apnea.

* **Correlation Matrix**: Stress levels showed negative correlations with sleep

quality and duration.

**4. Visualizations**

* **Histograms:** Displayed distributions for age, sleep duration, and stress levels.
* **Bar Charts:** Highlighted the prevalence of disorders across genders,

occupations, and BMI categories.

* **Correlation Heatmap:** Revealed relationships between features like stress,

physical activity, and sleep health.

**5. Key Insights**

1. Sleep disorders are influenced by gender, occupation, BMI, stress, and

physical activity.

1. Lower physical activity and higher stress levels are associated with poor

sleep quality.

1. High-stress occupations and irregular lifestyles significantly increase the

likelihood of sleep disorders.

**CHAPTER 5**

# Methodology

This section outlines the methodology used to develop the ML model for sleep disorder prediction, including tools, techniques, and implementation steps.

**5.1 Introduction to Languages (Front End and Back End)**

The project was implemented using **Python** for its robust libraries in data manipulation and machine learning. Future iterations may include a front-end interface for user interaction.

**5.2 Supporting Languages/Packages**

Key Python libraries used:

* **NumPy** and **Pandas** for data processing.
* **Matplotlib** and **Seaborn** for visualizations.
* **Scikit-learn** for preprocessing and ML algorithms.

**5.3 User Characteristics**

Designed for healthcare professionals, researchers, and individuals seeking an affordable diagnostic tool for sleep disorders.

**5.4 Constraints**

* Dataset size limited to 400 entries, impacting model generalizability.
* Model performance heavily relied on data quality and computational resources.

**5.5 Use Case Model/Flow Chart**

**The process flow includes:**

1. Data Collection and Preprocessing.
2. Feature Selection and Engineering.
3. Model Training (Decision Tree, Random Forest).
4. Prediction and Evaluation.
5. Visualization of Results.

**5.6 Database Design**

The dataset was structured as a CSV file with columns for demographics, health metrics, lifestyle factors, and sleep disorder type.

**5.7 Table Structure**

Key columns included:

* **Features**: Age, Gender, Sleep Duration, Stress Level, BMI, Daily Steps.
* **Target**: Sleep Disorder (None, Insomnia, Sleep Apnea).

**5.8 ER Diagrams**

While no ER diagram was explicitly created, data relationships between demographics, health metrics, and sleep disorders were conceptualized.

**5.9 Assumptions and Dependencies**

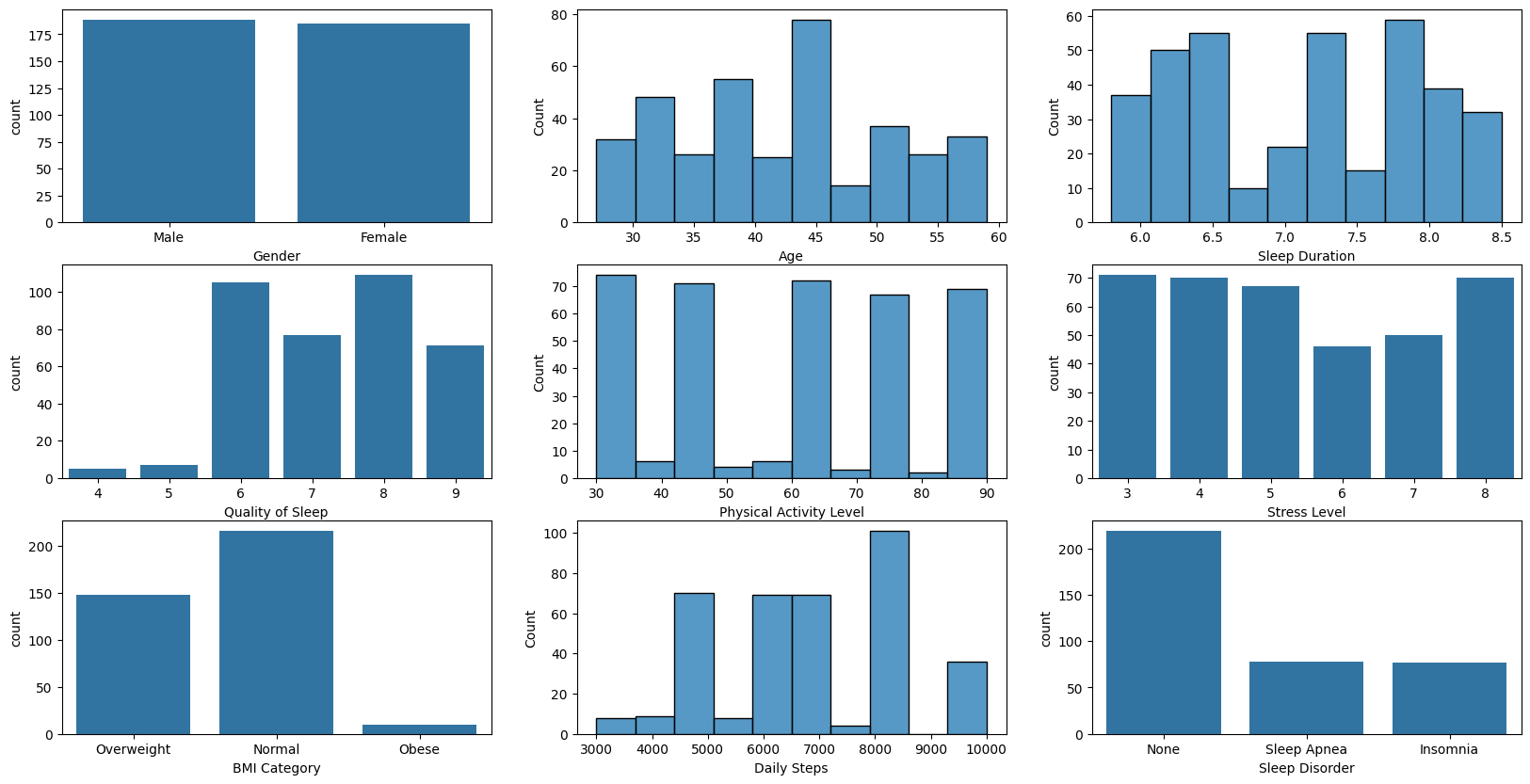
* Missing "Sleep Disorder" values were assumed to indicate "None."
* Data was considered accurate and free from major errors.
* Dependent on Python libraries for processing and modeling.

**5.10 ML Algorithm Discussion**

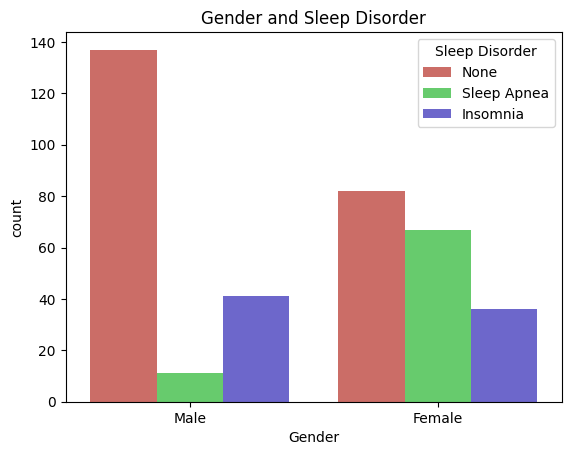
* **Decision Tree**: Achieved 87% accuracy, simple and interpretable.
* **Random Forest**: Outperformed Decision Tree with 89% accuracy, robust to noisy data.  
  Feature importance identified stress levels, BMI, and sleep duration as key predictors.

**5.11 Implementation with Screenshots/Figures**

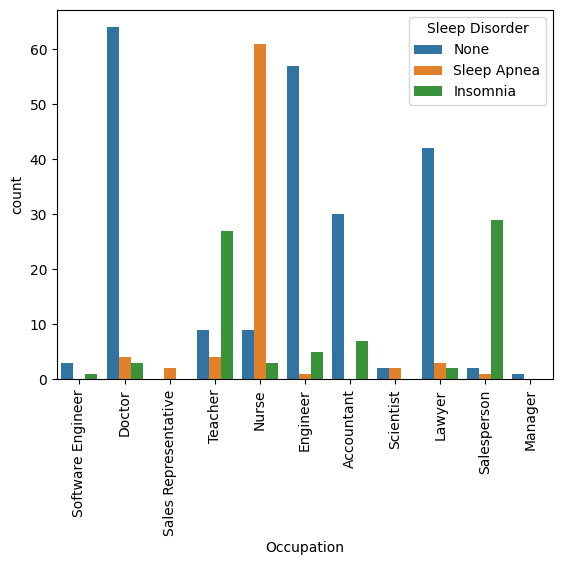
* **Steps**: Data preprocessing, feature encoding, model training, and evaluation.
* **Visuals**: Confusion matrices, correlation heatmap, and actual vs. predicted value distributions.



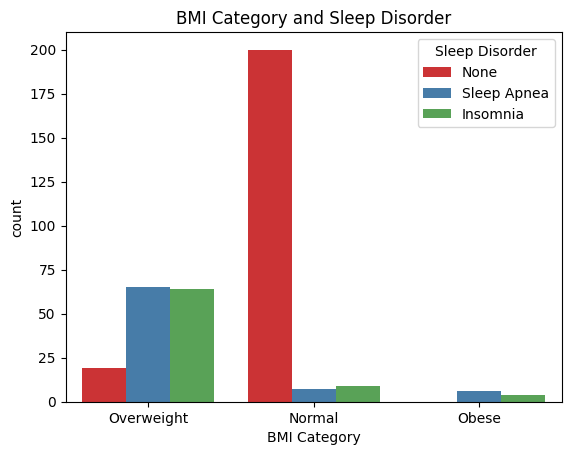
The number of males and females is almost equal, out of which majority of the people have age between 30-45 years. Most of the people have sleep quality greater than 5 which means there are getting sufficient sleep. Moreover, most of the people have normal BMI which directly relates with the distribution of sleep disorder which shows equal number of people with and without sleep disorder



Most of the males and females are not suffering from any sleep disorder. However females tend to have more sleep disorder as compared to males. The number of females suffering from Sleep Apnea is quite high as compared to males. But in contrast to that, greater number of males are suffering from Insomia as compared to females

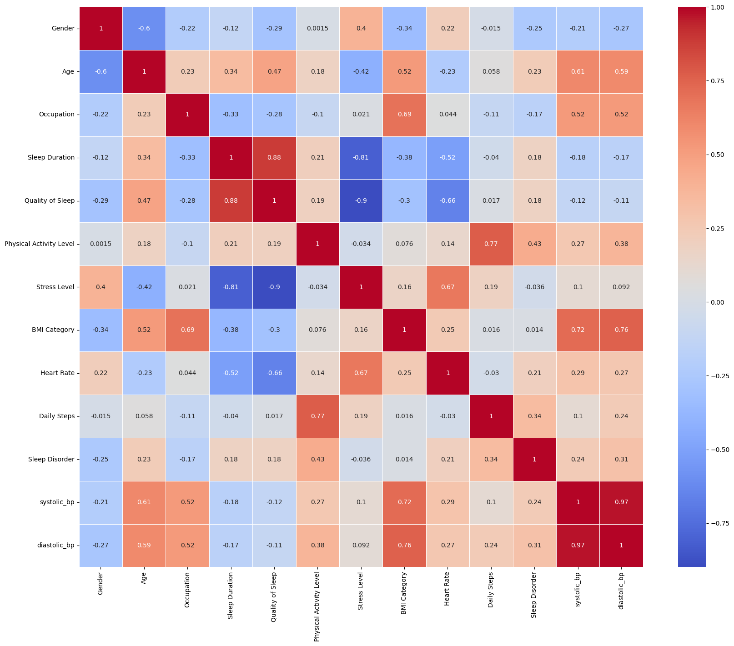


From the graph it is clear that the occupation has huge impact on the sleep disorder. Nurses are more subjected to have Sleep Apenea as compared to other occupations and very few of them have no sleep disorder. After nurses, the next most affected occupation is the Salesperson, which counts for the most suffering from Insomia followed by teachers. However there are some occupations where most of the people have very few instance of Sleep Apenea and Insomia such as Engineers, Doctors, Accountants, Lawyers. The Software ENgineers and Managers are so less in number so I cannot say much about that, But the occupation Sales Representative has shown only Sleep Apenea and no Insomia or no sleep disorder.



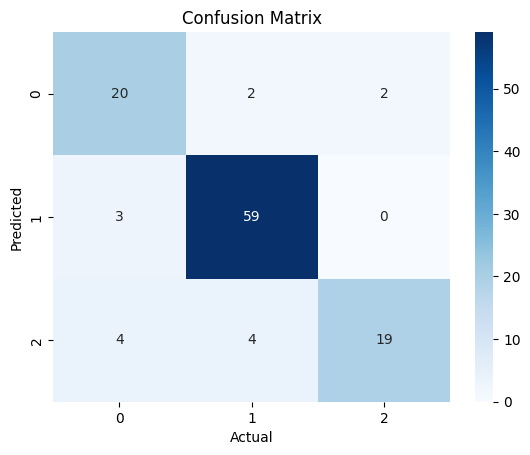
People with normal BMI are less likely to suffer from any sleep disorder. However, this is opposite in case of

Overweight and Obese people. Overweight are more likely to suffer more from sleep disordera than Obese people.

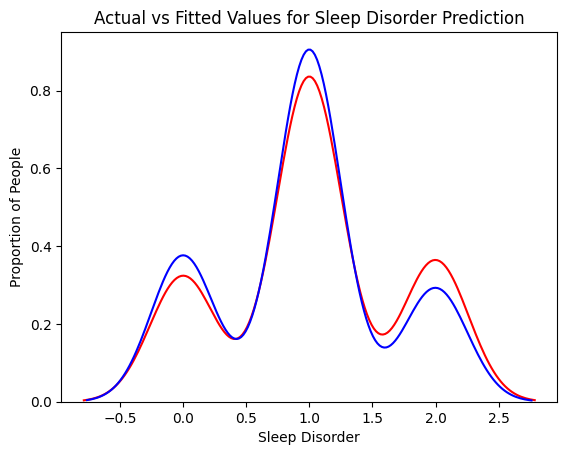


This heatmap shows how several lifestyle and health factors are correlated, providing information on possible

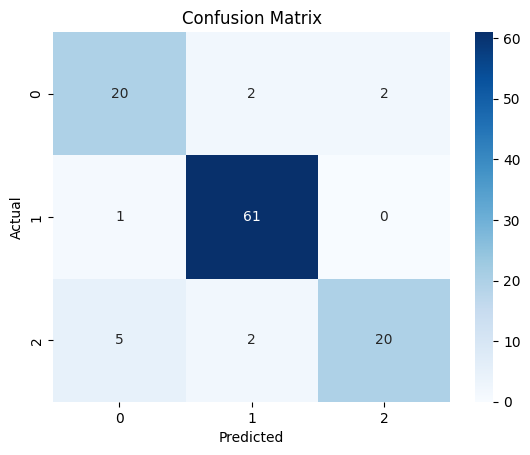
trends. For example, there is a substantial positive link between heart rate and stress levels, suggesting that elevated heart rates may result from increased stress. On the other hand, there is a negative correlation between BMI and physical activity levels, indicating that better weight management is facilitated by increased physical activity. Stress and sleep disturbances are positively correlated, suggesting that stress management may enhance the quality of sleep. Furthermore, age-related patterns like decreased physical activity and elevated BMI are noticeable, underscoring the significance of gradual lifestyle modifications. Effective management and prediction of health outcomes can be facilitated by these knowledge.



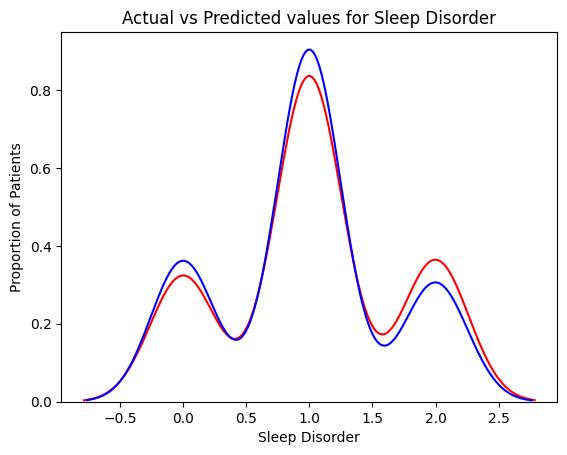
The diagonal boxes show the count of true positive results, i.e correct predictions made by the model. The off-diagonal boxes show the count of false positive results, i.e incorrect predictions made by the model.



The actual values are represented with red and the predicted ones with blue. As shown in the graph, the model's prediction are able to follow the curve of actual values but the predicted values are still different from actual ones. Therefore the model is not able to predict the values accurately.



The Random Forest Classifier model has greater accuracy than the Decision Tree Classifier model. The diagonal boxes count for the True Positives i.e correct predictions, whereas the off-diagonal boxes show the count of false positive results, i.e incorrect predictions made by the model. Since the number of false positve value is less, it shows that the model is good at predicting the correct results.



The Random forest classifier has improved accuracy as compared to the Decision Tree which is shown with the gap between the actual and predcited values which was wider incase of Descision Tree Classifier.

**CHAPTER 6**

# Results

The results of the project highlight the performance and insights gained from the machine

learning models used for sleep disorder prediction. The evaluation was based on metrics such as

accuracy, precision, recall, and F1 score, along with visual analysis through confusion matrices

and distribution plots.

**1. Model Performance**

Two supervised classification models, **Decision Tree** and **Random Forest**, were used. The

Random Forest model outperformed the Decision Tree in all key metrics:

| **Metric** | **Decision Tree** | **Random Forest** |
| --- | --- | --- |
| Accuracy | 87% | 89% |
| Precision (macro) | 0.85 | 0.87 |
| Recall (macro) | 0.83 | 0.85 |
| F1 Score (macro) | 0.84 | 0.86 |

* **Decision Tree**: Performed well with a training accuracy of 87%. However, it showed signs
* of overfitting, with slightly lower precision and recall on the test data.
* **Random Forest**: Demonstrated better generalization with higher accuracy and F1 score,

making it the preferred model for predicting sleep disorders.

**2. Confusion Matrix Analysis**

Confusion matrices were used to evaluate the classification accuracy for each class (None,

Insomnia, Sleep Apnea):

* **Random Forest** correctly classified most instances in all three categories, with minimal

false positives or negatives.

* Sleep Apnea cases had a slightly lower recall, indicating occasional misclassification as

Insomnia.

**3. Actual vs. Predicted Values**

Distribution plots comparing actual and predicted values showed a strong alignment for the

Random Forest model. While the Decision Tree predictions followed the actual values closely,

they displayed more variability.

**4. Feature Importance**

The Random Forest model provided insights into the significance of various features:

* **Stress Level**: The most influential factor in predicting sleep disorders.
* **BMI**: Strongly associated with Sleep Apnea, with overweight and obese categories

contributing significantly.

* **Sleep Duration**: A critical predictor, with insufficient sleep linked to both Insomnia and

Sleep Apnea.

* **Occupation**: Stressful jobs like nursing and sales showed higher disorder prevalence.

**5. Visualizations**

Key visualizations used to support findings included:

* **Confusion Matrix Heatmaps**: Clearly depicted true positives and false predictions for

each model.

* **Correlation Heatmap**: Highlighted the relationships between stress levels, physical

activity, and sleep disorders.

* **Bar Charts**: Showed the distribution of sleep disorders across gender, occupation, and
* BMI categories.

**6. Key Insights**

1. **Model Effectiveness**: Random Forest is better suited for this dataset due to its

robustness and superior performance metrics.

1. **Significant Predictors**: Stress levels, BMI, and sleep duration are critical in determining

sleep disorder likelihood.

1. **Disorder Trends**: Insomnia is more prevalent among males, while females show higher

instances of Sleep Apnea.

1. **Occupation Impact**: High-stress jobs like nursing and sales significantly increase the risk

of sleep disorders.

**CHAPTER 7**

**Conclusion and Future Scope**

## Conclusion

## The **Sleep Disorder Prediction** project demonstrated the effectiveness of machine

learning in identifying sleep disorders using demographic, lifestyle, and health factors.

The **Random Forest classifier**, with an accuracy of 89%, outperformed the Decision

Tree model and highlighted critical predictors like stress levels, BMI, and sleep

duration. The project offers a cost-effective, scalable alternative to traditional diagnostic

methods, providing valuable insights into sleep health trends such as higher Insomnia

rates among males and Sleep Apnea among females.

## Future Scope

**Dataset Expansion**: Use larger, more diverse datasets with additional features like comorbidities.

**Advanced Models**: Explore deep learning and hybrid models for improved accuracy.

**Real-Time Prediction**: Develop a user-friendly application for real-time data input and

predictions.

**Wearable Integration**: Incorporate data from smartwatches for continuous monitoring of sleep

patterns.

**Personalized Insights**: Extend the system to provide tailored sleep improvement

recommendations.

**Validation**: Conduct clinical trials to validate the model's performance in real-world settings.

**Broader Applications**: Adapt the model for fatigue monitoring or chronic illness management.

**BIBLIOGRAPHY**

1. Evangelopoulos, N., Zhang, X., & Prybutok, V. R. (2012). Latent Semantic Analysis: Five Methodological Recommendations. European Journal of Information Systems, 21(1), 70–86.

2. Canini, K. R., Shi, L., & Griffiths, T. L. (2009). Online Inference of Topics with Latent Dirichlet Allocation. Journal of Machine Learning Research, 5, 65–72.

3. Q. Mei & C. Zhai (2005). Discovering Evolutionary Theme Patterns from Text: An Exploration of Temporal Text Mining. The Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

4. J. L. Hurtado, A. Agarwal, & X. Zhu (2016). Topic Discovery and Future Trend Forecasting for Texts. Journal of Big Data, 3(7), 1–21.

5. Saini, S., Kasliwal, B., & Bhatia, S. (2013). Language Identification Using G-LDA. International Journal of Research in Engineering and Technology, 2(11), 42–45.

**APPENDIX**

The appendix provides supplementary information and supporting materials for the project,

offering insights into the dataset, preprocessing steps, and implementation details.

**1. Dataset Details**

* **Data Source**: A dataset of 400 entries with 13 features, including demographic (e.g., age,

gender), lifestyle (e.g., stress level, daily steps), and health-related variables (e.g., BMI,

sleep disorder).

* **Key Features**: Includes age (18–60 years), sleep duration, stress levels (1–10), BMI

categories, and sleep disorders (None, Insomnia, Sleep Apnea).

**2. Data Preprocessing**

* Missing values in "Sleep Disorder" were replaced with "None."
* Categorical variables were label-encoded.
* Blood pressure was split into systolic and diastolic components.

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

* **Visuals**: Correlation heatmaps, bar charts, and histograms highlight patterns between

stress levels, BMI, and sleep disorders.

**4. Algorithm Implementation**

* **Random Forest**: Trained with 100 estimators and optimized hyperparameters, achieving

the best accuracy.

* **Decision Tree**: Tuned to avoid overfitting but slightly less accurate.
* Code snippets for preprocessing, training, and evaluation are documented.

**5. Confusion Matrices and Metrics**

* Confusion matrices illustrate classification accuracy across disorder categories.
* Metrics such as precision, recall, and F1 scores validate model performance.

**6. Tools and Libraries**

* Python libraries: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn.
* Development environment: Jupyter Notebook.

**7. Limitations**

* Limited dataset size may affect generalizability.
* Imbalanced class distribution impacts predictions for less frequent disorders like Sleep

Apnea.