



SCHOOL OF COMPUTER SCIENCE

ITS61504 Data Mining

## INDIVIDUAL ASSIGNMENT

**HAND OUT DATE: FRIDAY, 29-May 2025**


**HAND IN DATE: 14<sup>th</sup> June 2025, 11:59 PM NPT**

**Instructions to students:**

- The assignment should be attempted individually
- Complete this cover sheet and attach it to your submission – this should be your first page.

**Module Learning Outcome:**

MLO1 - Demonstrate the ability to preprocess and explore datasets to uncover patterns and insights for a given scenario.

Student declaration:		
<i>I declare that:</i> <ul style="list-style-type: none"><li>▪ <i>I understand what is meant by plagiarism.</i></li><li>▪ <i>The implication of plagiarism and usage of AI generative tool have been explained to us by our lecturer.</i></li></ul> <i>This project is all our work, and I have acknowledged any use of the published or unpublished works of other people.</i>		
Name	Student ID	Signature
1. Rishika Maharjan	0371735	

## Table of Contents:

<b>1. Introduction.....</b>	<b>3</b>
1.1 Background of the case study.....	3
1.2 Problem Statement.....	4
1.3 Project Objectives.....	4
<b>2. Literature Review.....</b>	<b>5</b>
2.1 Summary of Related study.....	5
2.2 Proposed Methods and Techniques.....	6
2.3 Justification for the Method Selected.....	7
<b>3. Methodology.....</b>	<b>8</b>
3.1 Dataset Description and Source.....	8
3.2 Data Preprocessing Steps.....	9
3.3 Exploratory Data Analysis(EDA).....	11
3.4 Implementation of Data Mining Techniques.....	16
<b>4. Results and Interpretations.....</b>	<b>17</b>
4.1 Model Performance Evaluation.....	17
<b>5. Insights and Recommendation.....</b>	<b>21</b>
<b>6. Ethical Considerations.....</b>	<b>23</b>
6.1 Data Privacy and Confidentiality.....	23
6.2 Bias and Fairness.....	23
6.3 Transparency and Explainability.....	24
6.4 Responsible Usage.....	24
<b>7. Conclusion.....</b>	<b>24</b>
7.1 Summary of the Project and Findings.....	25
7.2 Challenges Faced and Limitations.....	25
7.3 Suggestions For Future Work.....	26
<b>8. References.....</b>	<b>26</b>

# **1. Introduction**

## **1.1 Background of the case study**

Sleep plays a crucial role in maintaining physical health, emotional well being and enhances the overall quality of life. However, in modern lifestyles people deal with high stress, irregular hectic work schedules, and inactive behavior have created a negative impact on quality of sleep across populations. Sleep disorders like insomnia and sleep apnea (a condition that affects your breathing while you're asleep) are increasingly common but frequently go undiagnosed or untreated. When we have sleep disturbances not only it causes issues in daily performance but also leads to various chronic health issues such as cardiovascular diseases, obesity, diabetes and mental health disorders. Early identification of individuals at risk of sleep disorder through data driven insights enables healthcare professionals to improve overall health outcomes.

In recent years, health and lifestyle data analysis through machine learning and data mining techniques has gained increased interest for predictive modeling. By analyzing patterns and correlations between lifestyle factors such as stress level, physical activity, BMI, and heart rate, we are enabled to develop models that anticipate the onset of sleep related problems.

## **1.2 Problem Statement**

The complexity and subjective nature of sleep diagnosis makes traditional diagnostic methods both time-consuming and dependent on patient self-report or expensive lab tests like polysomnography tests. This project addresses the possibility of whether an individual is likely to suffer from sleep disorder through analysis of self-reported lifestyle and health metrics. The primary focus is to utilize machine learning methods to detect sleep apnea or insomnia or to identify healthy sleep patterns in individuals.

### **1.3 Project Objectives**

To explore and understand the relationship between lifestyle habits and sleep disorders using real-world dataset.

To preprocess and transform the dataset for effective modeling, such as feature engineering and handling missing values.

To implement and compare multiple data mining classification techniques for predicting sleep disorder status.

In order to assess the performance of these models using standard classification metrics.

To deliver practical insights that can help inform healthcare decision making and encourage healthier lifestyle choices.

## **2. Literature Review**

### **2.1 Summary of Related study**

(Hidayat 2023) explored the Sleep Health and Lifestyle dataset and applied a Random Forest Classifier for predicting sleep disorders. The study included extensive preprocessing such as handling missing values, encoding categorical features and splitting the data. The model achieved high predictive accuracy and uncovered key important relationships particularly highlighting gender and stress patterns.

Alshammari et al.(2024) conducted a comparative study of multiple machine learning models such as KNN, SVM, Decision Tree, Random Forest, Naive Bayes, and Gradient Boosting on this dataset. Their results indicated that Gradient Boosting and Random Forest reported the highest accuracy up to 94.44% while SVM achieved an accuracy of 87.5% demonstrating the superior performance for ensemble and tree-based models.

A research paper by Shi investigated clinical data for severe obstructive sleep apnea(OSA) using six classifiers, including logistic regression, GBM, XGBoost, AdaBoost, Bagging, and MLP.

They achieved the best performance with GBM(AUC = 0.857) and highlighted the importance of variables like waist circumference and questionnaire scores via SHAP analysis.

Moreover, other lifestyle related studies have consistently identified obesity, heart rate, age, and gender as significant predictors of OSA and insomnia. These insights support the relevance of lifestyle features in predicting sleep disorders and guide the choice of models that can handle complex interactions.

## **2.2 Proposed Methods and Techniques**

For the prediction of sleep disorders, this study uses three well widely established machine learning classifiers: Random Forest, Decision Tree, and K-Nearest Neighbors(KNN). These methods are chosen because they are effective in handling both categorical and numerical data, as well as their proven success in healthcare related classification tasks.

### **1. Random Forest Classifier**

Random Forest Classifier is a widely used algorithm in the field of machine learning. This model ensembles learning method that creates multiple decision trees and combines their outputs to make final predictions. By averaging the outcomes of numerous trees, each trained on different random samples and features helps to mitigate overfitting(Bremian,2001). This model is particularly useful in handling noisy data and multicollinearity, which is common in health datasets with diverse lifestyle factors.

**Benefits:** High accuracy, resilience to overfitting, ability to handle mixed-type features, useful for ranking feature importance.

**Use Case:** Hidayat(2023) and Alshammri et al.(2024) showed that the Sleep Health dataset demonstrated Random Forest algorithm surpassed many baseline models by achieving an accuracy rate of over 90%.

## 2. Decision Tree Classifier

Decision Trees create a branch that leads to decision outcomes by recursively splitting the data based on the feature thresholds. This simple yet powerful approach makes trees easy to interpret, which is essential in health applications where stakeholders value transparency(Quinlan, 1996).

**Benefits:** Includes minimal data preprocessing, easy to visualize and understand, handles both categorical and numerical data.

**Use Case:** According to Alshammari et al.(2024), Decision Trees provided a high baseline accuracy on the Sleep Health dataset while also being interpretable and computationally efficient.

## 3. K-Nearest Neighbors(KNN)

The KNN algorithm is a non-parametric method that classifies a new data point based on the majority class of its '1' nearest neighbors in the feature space. It performs effectively with irregular decision boundaries and is especially beneficial for datasets of small to medium size(Cover & Hart, 1967).

**Benefits:** Works well with small datasets, no training time, easy to implement.

**Use Case:** In classification of sleep-related patterns, Alshammari et al.(2024) noted that KNN surpasses 85% accuracy in the observation, showing good performance compared to more complex models.

## 2.3 Justification for the Method Selected

The models were selected based on their data characteristics, model interpretability, and proven performance in comparable situations.

**Data Characteristics:** The datasets consist of both numerical(e.g, heart rate, steps) and categorical features(e.g, occupation, BMI category) variables. Random Forest and Decision Tree models are well suited for these types of mixed dataset without the need for extensive transformation.

**Model Interpretability:** Decision Trees and Random Forests provide model transparency, which is crucial in healthcare, where there should be logic to explain behind any predictions made to non experts stakeholders. KNN on the other hand offers simple predictions based on actual data points although it is less interpretable.

**Performance Benefits:** Random Forest is robust, performs feature selection internally, and scales effectively. The Decision Tree is quick and easy to read, making it great for understanding the data. KNN is a simple algorithm that can uncover the local patterns in data.

**Comparison to Other Models:** Although models like SVM and Naive Bayes are effective, they come with trade offs:

- SVM requires feature scaling and is less interpretable.
- Naive Bayes assumes that features are independent of one another, which may not be the case for complex lifestyle data.
- Although cited literature (Shi et al., 2023) verifies its high accuracy, Gradient Boosting was avoided to maintain simplicity and interpretability.
- Given these considerations, Random Forest, Decision Tree, and KNN were chosen for their balance of accuracy, interpretability, and compatibility with healthcare data types.

### 3. Methodology

#### 3.1 Dataset Description and Source

The dataset used in this project is titled "Sleep Health and Lifestyle Dataset", which is available on the kaggle platform. It consists of data of 374 individuals and includes variables related to their demographic characteristics, lifestyle habits, medical histories, and sleeping patterns. The aim of the dataset is to identify the factors that contribute to sleep disorders, like insomnia and sleep apnea.

The key features in the dataset include:

- **Demographic attributes:** Age, Gender, Occupation

- **Lifestyle Habits:** Daily Steps, Sleep Duration, Physical Activity Level
- **Medical Indicators:** Heart Rate, BMI, Blood Pressure
- **Sleep Disorder Level:** Sleep Disorder (Insomnia, Sleep Apnea, None)

**Description:**

Sleep Duration: Shows average hour of sleep per day

Physical Activity Level: Shows the number of the person engaged in any activity.

Blood Pressure: Indicates Systolic pressure over diastolic pressure

Heart Rate: Shows the heart beat rate per minute

BMI: Body Mass categorization

Columns and their Data Types :

- **Categorical:** Person ID, Gender, Occupation, Quality of Sleep, BMI Category, Sleep Disorder
- **Numerical :** Age, Sleep Duration, Stress Level, Blood Pressure, Heart Rate, Daily Steps

**Source :**

Kaggle: <https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset>

### 3.2 Data Preprocessing Steps

To ensure the quality and usability of the dataset, several preprocessing steps were taken:

**a. Data Cleaning**

A process where missing values are identified.



```
# Data Cleaning
# Handling missing values
print("\nin whole data the missing values are")
df.isnull().sum()
```

In whole data the missing values are	
Person ID	0
Gender	0
Age	0
Occupation	0
Sleep Duration	0
Quality of Sleep	0
Physical Activity Level	0
Stress Level	0
BMI Category	0
Blood Pressure	0
Heart Rate	0
Daily Steps	0
Sleep Disorder	219

Missing values were removed using `df.dropna()`.

```
# Removing the missing values
df_cleaned=df.dropna()
print("\nMissing values are removed")
```

Missing values are removed

Duplicate values were checked and none were found.

```
#Checking for duplicate variables
print(df.duplicated().sum())
```

0

The person Id column was dropped since it has no predictive value.

```
[8] # Drop unused column Person ID as it has no predictive value
df.drop("Person ID", axis=1, inplace=True)
```

## b. Data Transformation/ Encoding

Blood Pressure column from string format (e.g, “120/80”) was split into two numeric columns.

BMI category was ordinal encoded.

Categorical variables such as Gender, Occupation, BMI Category and Sleep Disorder were encoded to numerical values.

```

# Converting blood pressure into numerical value by splitting it into two parts
df[['BP_Systolic', 'BP_Diastolic']] = df['Blood Pressure'].str.split('/', expand=True).astype(int)
df = df.drop('Blood Pressure', axis=1)

# Ordinal Encdoing
bmi_order = ['Underweight', 'Normal', 'Overweight', 'Obese']
df['BMI Category'] = pd.Categorical(df['BMI Category'], categories=bmi_order, ordered=True).codes

# Encoding categorical variables
categorical_cols = df.select_dtypes(include='object').columns
le = LabelEncoder()
for col in categorical_cols:
    df[col] = le.fit_transform(df[col])

```

### c. Feature Engineering

New features like Stress\_Sleep\_Interaction and Activity\_Sleep likely improved the model by combining lifestyle factors. Age binning helped models generalize better.

Interaction terms were created:

Stress\_Sleep\_Interaction = Stress Level \* Sleep Quality

```

#
df['Stress_Sleep_Interaction'] = df['Stress Level'] * df['Quality of Sleep']
df['Activity_Sleep'] = df['Physical Activity Level'] * df['Sleep Duration']

# Age Binning grouping the age groups into different categories
df['AgeGroup'] = pd.cut(df['Age'], bins=[0, 18, 30, 50, 100], labels=['Teen', 'Young', 'Adult', 'Senior'])
df['AgeGroup'] = le.fit_transform(df['AgeGroup'])

```

### d. Feature Scaling

```

# Seperate Features and target
X = df.drop('Sleep Disorder', axis=1)
y = df['Sleep Disorder']

#Standarize featurues
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

```

### e. Feature Selection

```

#Random Forest for feature importance
rf_selector = RandomForestClassifier(random_state=42)
rf_selector.fit(X_scaled, y)

#Get and sort feature
importances = pd.Series(rf_selector.feature_importances_, index=X.columns)
sorted_importances = importances.sort_values(ascending=False)

importances_df = pd.DataFrame({
    'feature': sorted_importances.index,
    'Importance': sorted_importances.values
})

#Select top N Features (e.g, top 10)
top_features = importances.head(5).index.tolist()
plt.figure(figsize=(8, 5))
sns.barplot(x='Importance', y='Feature', data=importances_df.head(10))
plt.title('Top Feature Importances from Random Forest')
plt.xlabel('Importance Score')
plt.ylabel('Feature')
plt.tight_layout()
plt.show()

```

### 3.3 Exploratory Data Analysis(EDA)

EDA is a critical step used to understand the structure, patterns, and relationships in the data before modeling.

#### Data exploration:

All the information of data is displayed using a certain method before doing anything to get the data information.

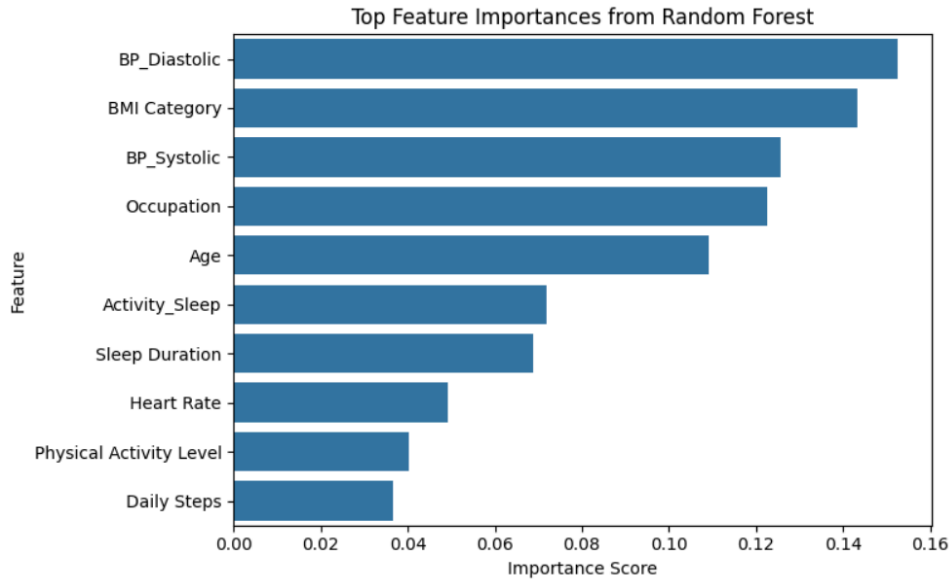
```
# Checking information about the data
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 374 entries, 0 to 373
Data columns (total 13 columns):
#   Column              Non-Null Count  Dtype  
---  -
0   Person ID           374 non-null   int64  
1   Gender              374 non-null   object  
2   Age                 374 non-null   int64  
3   Occupation          374 non-null   object  
4   Sleep Duration      374 non-null   float64 
5   Quality of Sleep    374 non-null   int64  
6   Physical Activity Level 374 non-null   int64  
7   Stress Level        374 non-null   int64  
8   BMI Category        374 non-null   object  
9   Blood Pressure      374 non-null   object  
10  Heart Rate          374 non-null   int64  
11  Daily Steps         374 non-null   int64  
12  Sleep Disorder       155 non-null   object  
dtypes: float64(1), int64(7), object(5)
memory usage: 38.1+ KB
```

#### Visualization

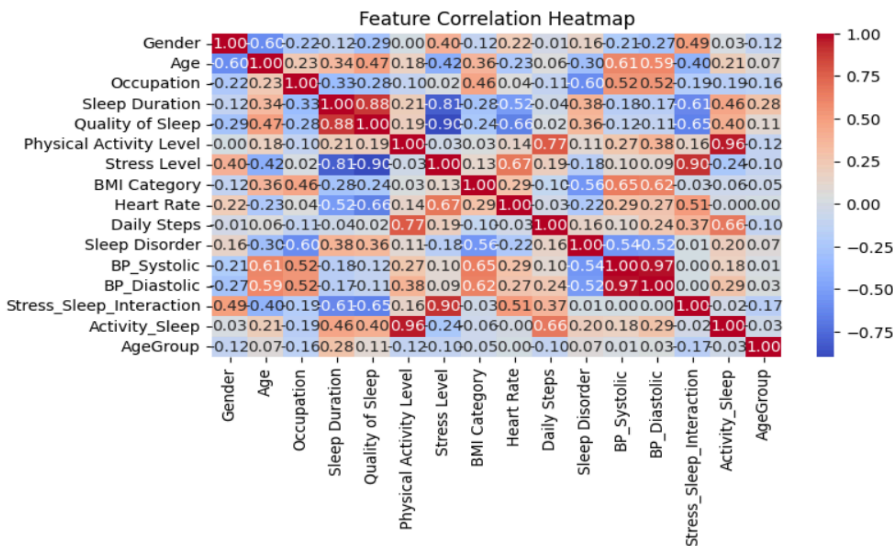
##### a. Feature Selection using Random Forest

Top required features are ranked in this part before moving forward to make the model more accurate.



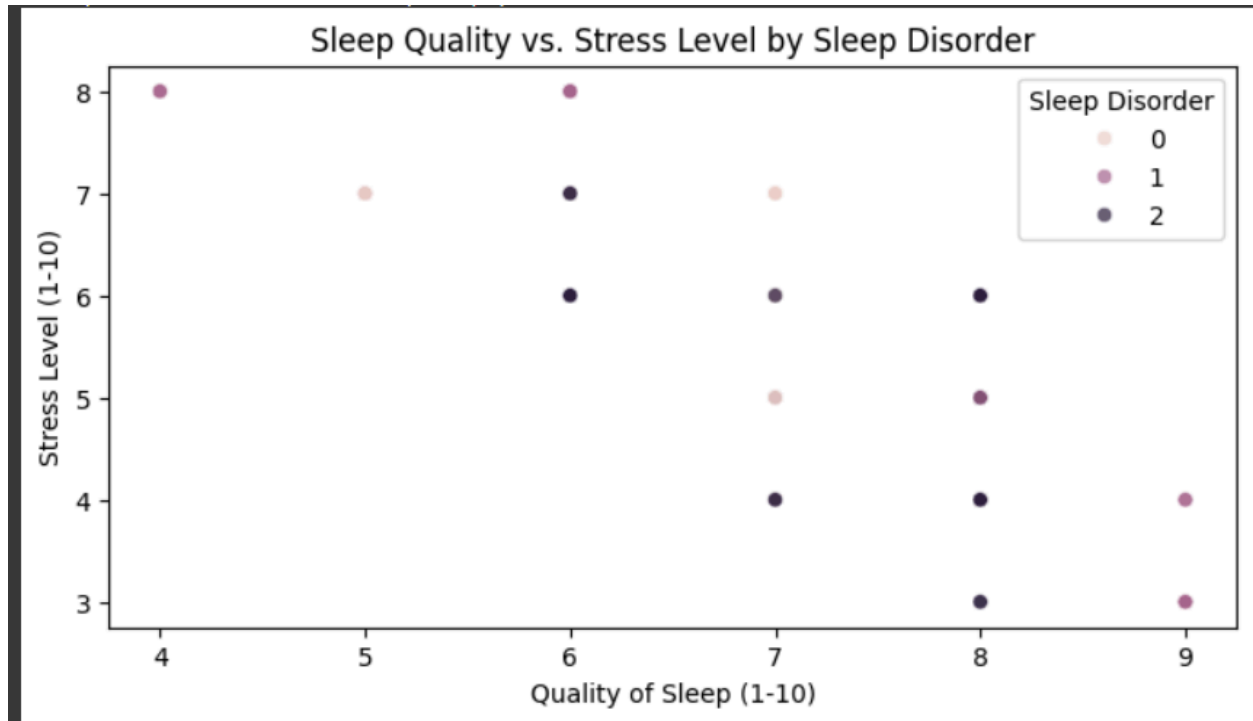
## b. Correlation Heatmap

The correlation heatmap in my code shows the correlation between numerical variables. It helps to identify which features may be redundant or predictive.



### c. ScatterPlot

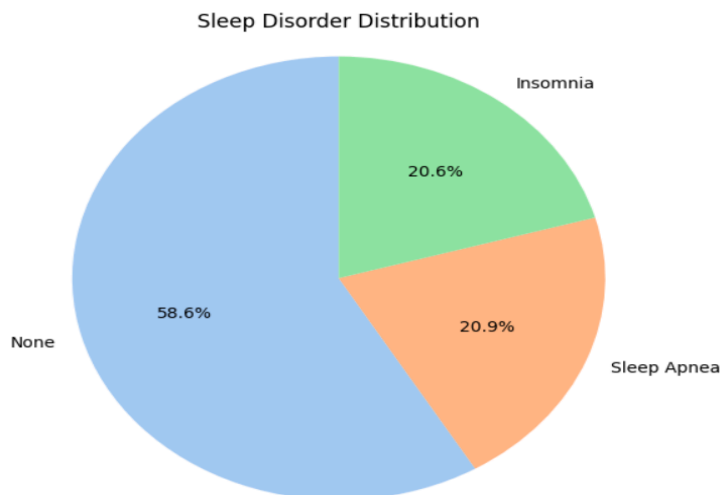
It helps to visualize how stress level and sleep quality relate to different sleep disorder types.



### d. Distribution Analysis

#### A. Pie Chart

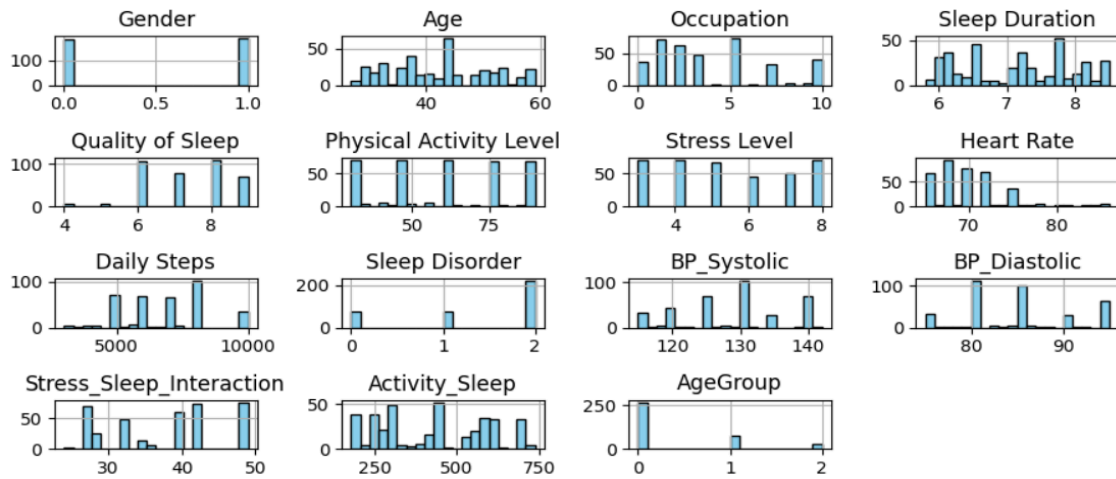
The pie chart shows the distribution of the target variables None, Sleep Apnea and Insomnia.



### e. Histogram of Numeric Variables

The histogram shows the distribution of each numerical variable and helps to spot skewed distributions.

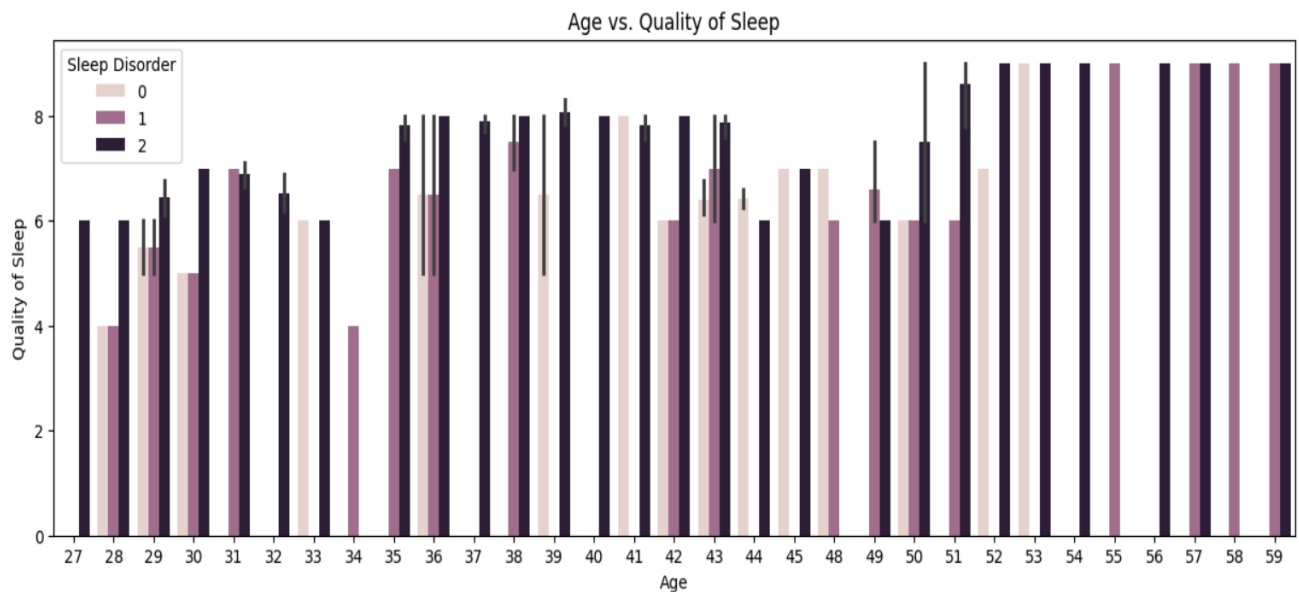
Histograms of Numerical Features



Some Predictions made to visualize how sleep factor is affected by lifestyles in their daily life

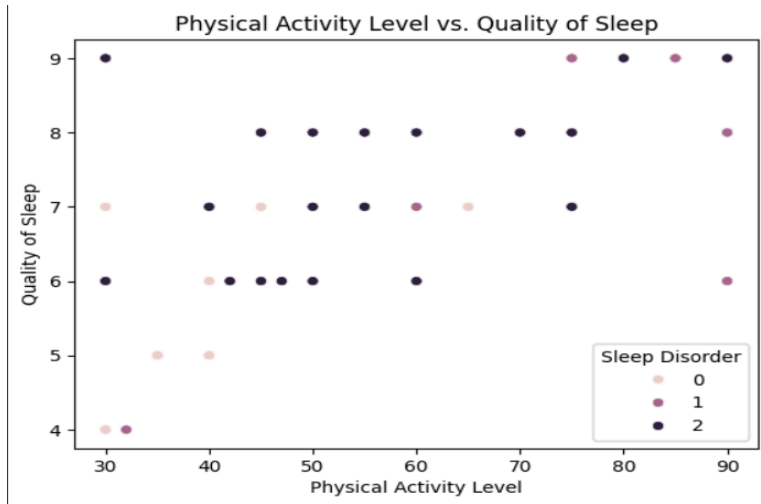
### f. Barplot

It shows the quality of sleep or sleep disorder factor a person has according to the age of that individual.



### g. Scatterplot

Predicting if an increased physical activity results to quality of sleep



## 3.4 Implementation of Data Mining Techniques

Here, I have applied three classifiers for multi-class classification to predict the type of Sleep disorder.

### 1. Random Forest Classifier

Random Forest is a method that builds multiple decision trees and combines their results to improve accuracy and reduce overfitting. This model can handle large feature sets, interactions and non linear patterns.

### 2. Decision Tree Classifier

Decision Tree is a simple interpretable model that splits data based on feature values to make predictions. It can handle missing data and helps visualize the decision paths.

### 3. K-Nearest Neighbors Classifier(KNN)

A distance based algorithm that classifies a data point based on the majority class among its k nearest neighbors. No training is required and is simple with adaptability features.

#### 4. Model Evaluation:

To evaluate the model performance, I used the following metrics:

- **Accuracy:** Proportion of correctly classified instances among all the predictions made.
- **Precision:** Out of all predicted ones, how many were actually correctly predicted.
- **Recall:** Out of all positive instances, how many were correctly identified.
- **F1-score:** Balance means precision and recall.
- **Confusion Matrix:** Visualized for each classifier to show prediction distribution across the classes.

## 4. Results and Interpretations

### 4.1 Model Performance Evaluation

#### Train Test and split

```
x = df.drop('Sleep Disorder', axis=1)
y = df['Sleep Disorder']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
scaler = StandardScaler()
print("Preprocessing complete. Shapes:")
print(f"X_train: {X_train.shape}, y_train: {y_train.shape}")
print(f"X_test: {X_test.shape}, y_test: {y_test.shape}")
```

Preprocessing complete. Shapes:  
X\_train: (299, 15), y\_train: (299,)  
X\_test: (75, 15), y\_test: (75,)

#### SMOTE

Using smote to balance the target variable for better accuracy results.



```

➡ Before SMOTE: Sleep Disorder
2    176
1     62
0     61
Name: count, dtype: int64
After SMOTE: Sleep Disorder
0    176
2    176
1    176
Name: count, dtype: int64

```

The three Classification models used for predicting sleep disorders were:

- **Random Forest**

```

➡ RandomForestClassifier: 0.88
      precision    recall  f1-score   support

      None         0.72      0.81      0.76         16
      Insomnia      0.85      0.69      0.76         16
      Sleep Apnea    0.95      0.98      0.97         43

   accuracy          0.88          75
  macro avg         0.84      0.83      0.83          75
 weighted avg         0.88      0.88      0.88          75

```

- **Decision Tree**

```

➡ Decision Tree Accuracy: 0.88
      precision    recall  f1-score   support

      None         0.81      0.81      0.81         16
      Insomnia      0.76      0.81      0.79         16
      Sleep Apnea    0.95      0.93      0.94         43

   accuracy          0.88          75
  macro avg         0.84      0.85      0.85          75
 weighted avg         0.88      0.88      0.88          75

```

- **K-Nearest Neighbors(KNN)**

KNeighborsClassifier: 0.88					
	precision	recall	f1-score	support	
None	0.72	0.81	0.76	16	
Insomnia	0.85	0.69	0.76	16	
Sleep Apnea	0.95	0.98	0.97	43	
accuracy			0.88	75	
macro avg	0.84	0.83	0.83	75	
weighted avg	0.88	0.88	0.88	75	

Here each model was evaluated using:

- Accuracy
- Precision
- Recall
- F1- Score
- Confusion Matrix

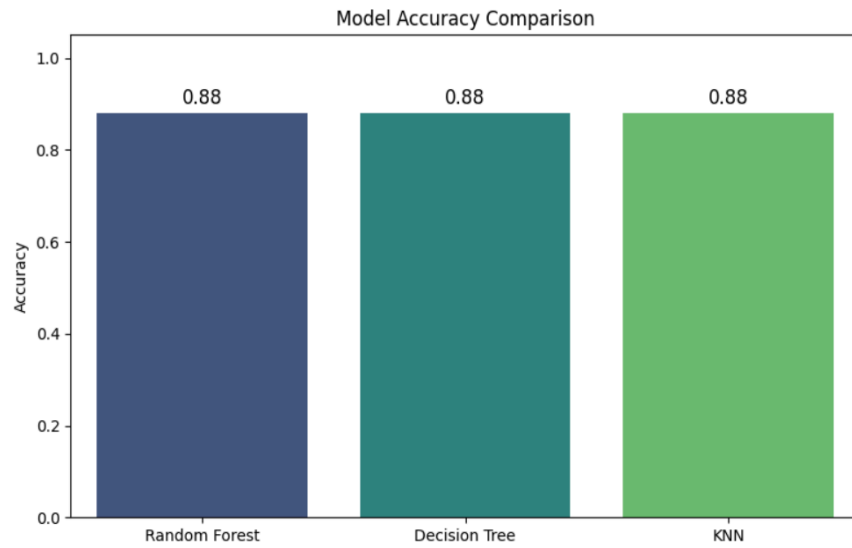
```
def evaluate_model(name, y_true, y_pred):
    print(f"----- {name} Evaluation -----")
    print("Accuracy:", accuracy_score(y_true, y_pred))
    print("Precision:", precision_score(y_true, y_pred, average='weighted'))
    print("Recall:", recall_score(y_true, y_pred, average='weighted'))
    print("F1 Score:", f1_score(y_true, y_pred, average='weighted'))
    print()
```

## 4.2 Visualizations Used for Insights

### Model Accuracy Comparison Using the three trained Models

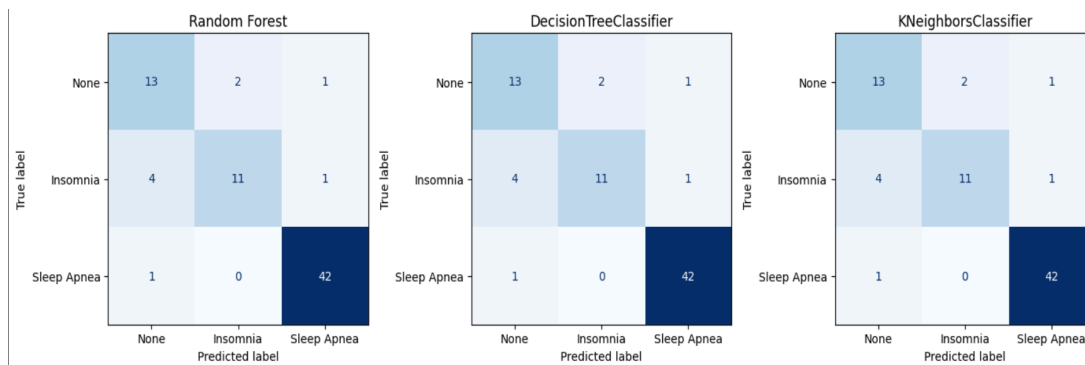
Here, the performance of Random Forest, Decision Tree, and KNN all had similar high accuracy that is 88%.

### Using Barplot

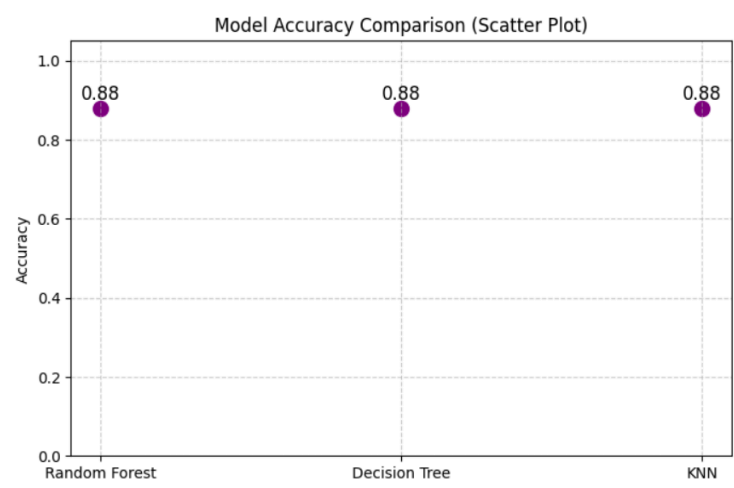


## Using Confusion Matrix

Confusion Matrix displays the class wise prediction of strengths and weaknesses for each model.



## Using Scatter Plot



## 5. Insights and Recommendation

### 5.1 Insights from Exploratory Data Analysis(EDA):

- **Stress and Sleep Quality Correlation:**

A negative correlation was observed between stress levels and sleep quality where individuals with higher stress generally have poor sleep quality.

- **Physical Activity Improves Sleep:**

Individuals with higher physical activity levels tend to have better quality of sleep, especially those without a sleep disorder.

- **Age Factor:**

A pattern where middle-aged and older individuals tend to have more sleep disorders, particularly sleep apnea.

Also, sleep quality decreases with age due to health complications.

- **Occupation Influence:**

People having occupations with sedentary lifestyles(e.g, software engineers, managers) showed a higher occurrence of insomnia in the visualization chart.

- **Sleep Disorder Type Distribution:**

In the dataset, around 59% of individuals have diagnosed sleep disorder(mostly sleep apnea, followed by Insomnia), while the rest report no disorder.

## **5.2 Model Insights From Evaluation**

- All three models performed equally well, whereas Random Forest slightly performs better at handling imbalance classes after applying SMOTE.
- Among the sleep disorder analysis, sleep apnea was classified most accurately, with precision and recall above 95% in Random Forest and KNN.
- Insomnia detection had lower recall suggesting it may be harder to detect one present.

## **5.3 Actionable Recommendation**

### **For individuals:**

- Find a way for stress management by participating in activities like yoga, mindfulness, or counseling as a part of therapy.
- Physical activity should be done more often as it could enhance the sleep quality especially for sedentary individuals.
- People with high BMI, elevated blood pressure, or who are of an older age group should be screened routinely.
- 

### **For Decision Makers in Health Institutions:**

- Using this model as a part of a digital health screening system to pre-identify individuals at risk of sleep disorders.
- Data driven awareness programs targeting age groups or professions most affected.

- Introduce work life balance policies for stress reduction in high risk job categories.

#### **For Model Improvement:**

- Collect more balanced data for disorders like insomnia.
- Include other variables such as screen time, caffeine/alcohol intake, or mental health for improved prediction.

## **6. Ethical Considerations**

### **6.1 Data Privacy and Confidentiality**

#### **Sensitive Data:**

The dataset related to the health industry more often includes personal health related attributes such as sleep habits, stress level, and physical activity.

#### **Protection Measures:**

Although the data sometimes may be anonymized, care must be taken to avoid the re-identification of individuals. If the model is to be used in real world application data protection regulations must be used to ensure confidentiality.

#### **Recommendation:**

If any deployment is to be made make sure to implement data encryption, access control, and consent mechanisms when collecting or processing personal data.

### **6.2 Bias and Fairness**

#### **Potential Biases:**

- ❖ Class imbalance was observed for example fewer cases of “Insomnia compared to “Sleep Apnea” which could lead to bias model prediction.

- ❖ SMOTE was used to balance the classes during training, which helps to lower the prediction bias.

### **Fairness Risks:**

Unfair outcomes can occur due to some demographic groups(e.g, gender, age,occupation) may be over-or under-represented.

### **Mitigation Strategy:**

- ❖ Monitor model fairness across subgroups.
- ❖ Use evaluation metrics for fairness.
- ❖ Retrain using balanced, diverse datasets.

## **6.3 Transparency and Explainability**

Models like Random Forest or Decision Trees offer certain explainability via feature importance. However, users(patients, doctors,etc) must understand how predictions are made to build trust.

## **6.4 Responsible Usage**

### **Use-case boundaries:**

The model is not a diagnostic tool but a screening or awareness aid so it should only support, not replace clinical judgement.

### **Deployment Risk:**

Misclassification could lead to over-treatment or negligence.

### **Guideline:**

- ❖ Always accompany prediction with confidence levels.
- ❖ Design system to alert a clinician, not act independently.

## **7. Conclusion**

### **7.1 Summary of the Project and Findings**

The goal of this project was to build a predictive model to determine whether a person is likely to suffer from sleep disorder, such as insomnia, sleep apnea, or none at all based on their lifestyle and health related attributes.

Thorough data preprocessing and transformation were performed, such as handling missing values, converting categorical data, and normalizing features.

To balance the class distribution and reduce model bias, SMOTE was employed.

Three classification models Decision Tree, Logistic Regression, and KNN were built and compared according to their performances.

All three models had the same accuracy but the decision tree had slightly high accuracy and well balanced precision-recall metrics.

Important predictors included Occupation, BMI Category, Physical Activity Level, and Stress Level.

### **7.2 Challenges Faced and Limitations**

#### **Class Imbalance :**

Originally, it was challenging to train balanced models due to the fewer instances of insomnia and sleep apnea. Although synthetic oversampling is beneficial, it may not be able to completely replicate real distributions.

#### **Limited Dataset Size:**

The dataset was relatively small making the model prone to overfitting and restricting its generality.

#### **Lack of Time-Series Data:**

Sleep patterns are time dependent, and static features miss important time-based trends.



**Future Subjectivity:**

Factors such as sleep duration or stress level were probably reported by individuals which could have led to inconsistent or biased results.

**Model Explainability:**

Certain classification models were accurate but lacked transparency for non technical users such as health workers.

**7.3 Suggestions For Future Work****Expand the Dataset:**

Include a large number of participants from diverse demographic groups and health conditions. Gather time series sleep tracking data(e.g, from wearables) for better modeling.

**Feature Enhancement:**

Include new variables such as caffeine intake, screen time, mental health indicators, and sleep quality scores. For higher accuracy also include clinical data such as sleep study findings.

**Model Improvement:**

Use ensemble models such as Gradient Boosting or XGBoost for better accuracy. Apply Explainable AI(XAI) methods to enhance transparency and trust.

**Real-World Integration:**

Make it available as a convenient mobile app or dashboard tool for sleep clinics to help doctors in screening patients more effectively. Based on model outputs, provide users with tailored lifestyle suggestions.

## 8. References

- Atiqur, Rahman. "Improving Sleep Disorder Diagnosis Through Optimized Machine Learning Approaches." IEEE Access, vol. 13, 2025, pp. 20989 - 21004, <https://ieeexplore.ieee.org/document/10856004>. Accessed Friday June 2025.
- Hidayat, Idfian Azhar. "Classification of Sleep Disorders Using Random Forest on Sleep Health and Lifestyle Dataset." Journal of Dinda Data Science, Information Technology, and Data Analytics, vol. 3 No. 2, no. 2023, 2023, p. 76. [rResearchgate.net/](https://www.researchgate.net/), <https://www.researchgate.net/>. Accessed Friday June 2025.
- Talal, Alshammari. "Applying Machine Learning Algorithms for the Classification of Sleep Disorders." no. 2024, 2024, p. 99. <https://www.researchgate.net/>, [https://www.researchgate.net/publication/378816773\\_Applying\\_Machine\\_Learning\\_Algorithms\\_for\\_the\\_Classification\\_of\\_Sleep\\_Disorders](https://www.researchgate.net/publication/378816773_Applying_Machine_Learning_Algorithms_for_the_Classification_of_Sleep_Disorders). Accessed Friday June 2025.
- T, Cover, and Hart P. "Nearest neighbor pattern classification." vol. 13, no. 1, 1967, pp. 21-27. IEEE Transactions on Information Theory, <https://ieeexplore.ieee.org/document/1053964/authors#authors>. Accessed Friday June 2025.
- Links:
- <https://ieeexplore.ieee.org/document/10856004>
- [https://www.researchgate.net/publication/372975033\\_Classification\\_of\\_Sleep\\_Disorders\\_Using\\_Random\\_Forest\\_on\\_Sleep\\_Health\\_and\\_Lifestyle\\_Dataset](https://www.researchgate.net/publication/372975033_Classification_of_Sleep_Disorders_Using_Random_Forest_on_Sleep_Health_and_Lifestyle_Dataset)
- [https://www.researchgate.net/publication/378816773\\_Applying\\_Machine\\_Learning\\_Algorithms\\_for\\_the\\_Classification\\_of\\_Sleep\\_Disorders](https://www.researchgate.net/publication/378816773_Applying_Machine_Learning_Algorithms_for_the_Classification_of_Sleep_Disorders)
- <https://ieeexplore.ieee.org/document/1053964>

