REVA UNIVERSITY
Bengaluru, India

SCHOOL OF COMPUTING AND INFORMATION TECHNOLOGY

SUBJECT - MACHINE LEARNING

SLEEP HEALTH AND LIFESTYLE ANALYSIS

SUBMITTED BY

NAME: CHAITRA LAKSHMI VARDHAN . C
SRN: R22EK033
DATE OF SUBMISSION: 10/10/2024
SIGNATURE:

VERIFIED BY

MARKS:

SIGNATURE OF FACULTY:

SLEEP HEALTH AND LIFESTYLE ANALYSIS

Introduction

This dataset explores the relationship between sleep health and everyday lifestyle choices. It looks at different factors like age, gender, and occupation, as well as important health details such as how many hours people sleep, the quality of their sleep, their physical activity, stress levels, and body weight categories. The aim is to find out how these lifestyle factors affect sleep quality and duration. By understanding these connections, we hope to discover ways to improve sleep and overall health, helping people lead happier and healthier lives.

Dataset Overview:

- **Number of samples (rows):** 1374
- **Number of features (columns):** 13

Column names:

1. **Person ID:** Unique identification for each individual.
2. **Gender:** Gender of the individual (e.g., Male, Female).
3. **Age:** Age of the individual.
4. **Occupation:** Occupation category of the individual (e.g., student, employed).
5. **Sleep Duration:** Total hours of sleep per day.
6. **Quality of Sleep:** Quality assessment of sleep on a scale of 1-9.
7. **Physical Activity Level:** Daily physical activity level on a scale from 0-99.
8. **Stress Level:** Stress level measured on a scale from 1-9.
9. **BMI Category:** BMI classification (e.g., underweight, normal, overweight, obese).
10. **Blood Pressure:** Blood pressure readings (e.g., systolic/diastolic).
11. **Heart Rate:** Average heart rate during the day.
12. **Daily Steps:** Number of steps taken daily.13. **Sleep Disorder:** Type of sleep disorder if applicable (e.g., Insomnia, Sleep Apnea).

	Person ID	Gender	Age	Occupation	Sleep Duration	\
0	1	Male	27	Software Engineer	6.1	
1	2	Male	28	Doctor	6.2	
2	3	Male	28	Doctor	6.2	
3	4	Male	28	Sales Representative	5.9	
4	5	Male	28	Sales Representative	5.9	

	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	\
0	6	42	6	Overweight	
1	6	60	8	Normal	
2	6	60	8	Normal	
3	4	30	8	Obese	
4	4	30	8	Obese	

	Blood Pressure	Heart Rate	Daily Steps
0	126/83	77	4200
1	125/80	75	10000
2	125/80	75	10000
3	140/90	85	3000
4	140/90	85	3000

The **Sleep Health and Lifestyle Dataset** is classified as **structured data** because it is organized in a tabular format with clearly defined rows and columns. Each column represents a specific attribute, such as Person ID, Gender, Age, Sleep Duration, and Quality of Sleep, while each row corresponds to an individual record. This structured organization makes it easy for querying, analysis, and for various data manipulation techniques.
The **sleep health and lifestyle** dataset can be accessed from multiple online forums like **Kaggle**, making it well-known **public** resource.

The size of the dataset is determined by the number of rows and number of columns before performing PCA. In this dataset it has **1374 rows** and **13 columns.** It means that there are 1374 individual records, each described by 12 attributes.

```
Number of rows (samples): 1374
Number of columns (features): 12
```

The labels in the **Sleep Health and Lifestyle Dataset** are found in several categorical columns, such as Gender, Quality of Sleep, and Physical Activity Level. These labels are accurate and meaningful, directly representing key attributes related to health and lifestyle.

The following code can be used to identify missing values :

```
import pandas as pd
data = pd.read_csv('sleepanalysis.csv')
missing_values = data.isnull().sum()
print("Missing values in each column:")
print(missing_values)
print("\nData types of each column:")
print(data.dtypes)
print("\nChecking for negative values in numeric columns:")
for column in ['Age', 'Sleep Duration', 'Physical Activity Level', 'Stress Level', 'Blood Pre
    if (data[column] < 0).any():
        print(f"Negative values found in {column}")
    else:
        print(f"No negative values in {column}")
```

The output is as follows :

```
Missing values in each column:
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
dtype: int64
```

ALGORITHM : 1
PCA (PRINCIPAL COMPONENT ANALYSIS)

The objective of applying Principal Component Analysis (PCA) to this sleep analysis dataset is to reduce the number of features while retaining as much significant information as possible. Given the multiple variables—such as demographic details, sleep patterns, physical activity levels, and health metrics—PCA can simplify the dataset by combining highly correlated features and uncovering underlying patterns.

PCA is particularly valuable here for several reasons:

- **Dimensionality Reduction:** It reduces the complexity of the dataset by consolidating features, making the data more manageable for analysis.

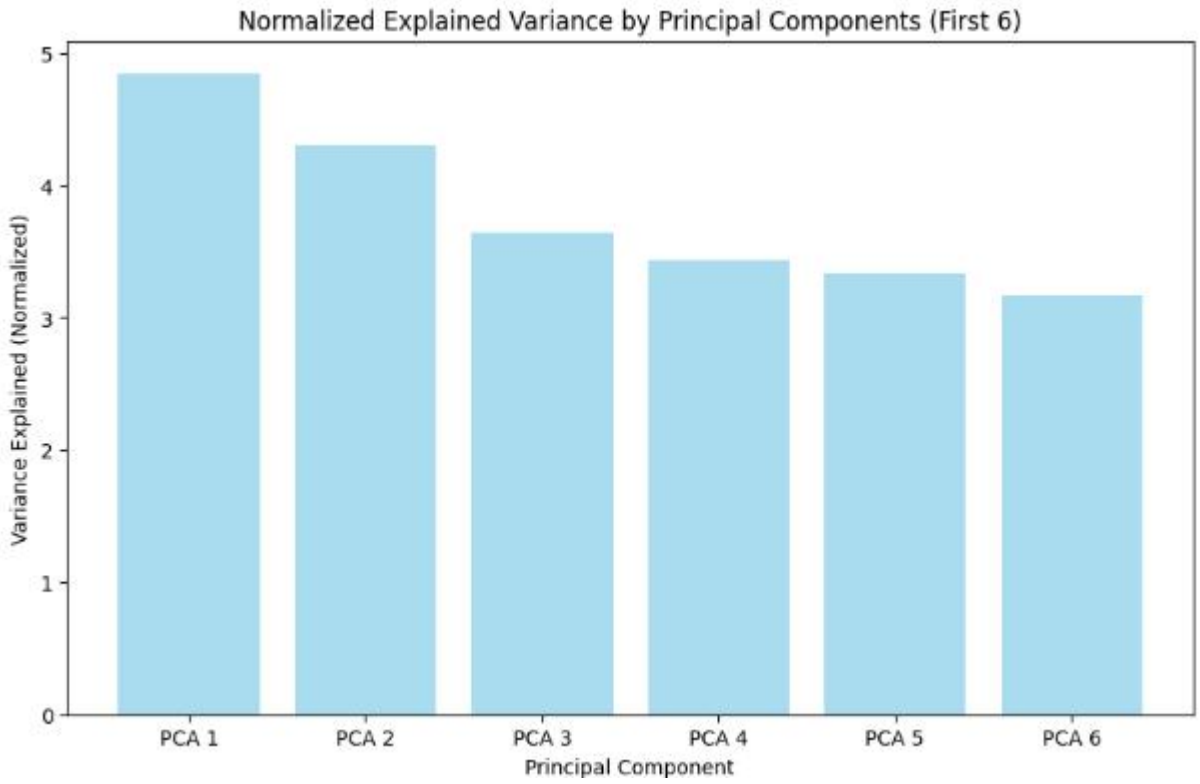
- **Highlighting Patterns**: By identifying and combining correlated attributes, PCA reveals significant trends that might otherwise be less visible.

By applying PCA, we aim to improve computational efficiency and reduce multidisciplinary issues, which can occur when some features are highly interdependent. Additionally, PCA helps visualize the dataset in fewer dimensions, allowing us to detect patterns or clusters within the data, such as groupings based on similar sleep quality, stress levels, or physical activity. This streamlined analysis makes it easier to interpret the relationships between sleep health and lifestyle factors.

```
import pandas as pd
categorical_columns = ['Gender', 'Occupation', 'BMI Category']
df_encoded = pd.get_dummies(df, columns=categorical_columns, drop_first=True)
```

Applying PCA on the dataset :

```
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)
```

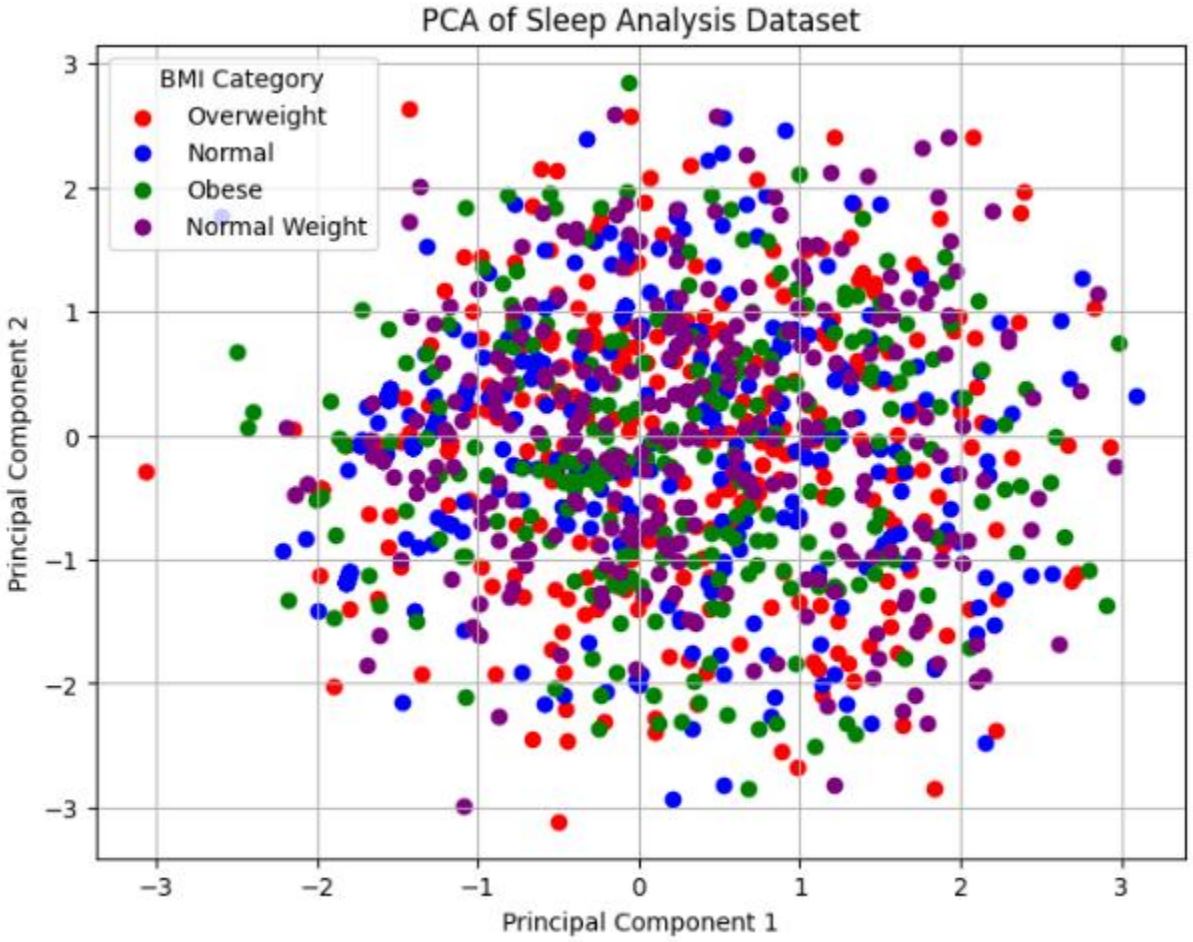


Here, the explained variance ratio for each principal component is as follows

Principal Component	Variance Explained (%)
PCA 1	4.85
PCA 2	4.30
PCA 3	3.64
PCA 4	3.44
PCA 5	3.34
PCA 6	3.17

The analysis of the principal components shows that Principal Component 1 (PC1) explains the largest portion of the variance at 4.85%, indicating that it is the most significant factor in differentiating the songs in the dataset. Following PC1, Principal Component 2 (PC2) accounts for 4.30% of the variance. Subsequent components, including PC3 (3.64%), PC4 (3.44%), PC5 (3.34%), and PC6 (3.17%), contribute progressively smaller amounts of variance.

The explained variance ratios reveal that PC1, while the most significant, only captures 4.85% of the variance. This suggests that no single component overwhelmingly influences the dataset, and a broader set of components may be needed to capture the full diversity of the data. However, PC1's slight dominance indicates it is still a crucial component, potentially reflecting primary characteristics.



ALGORITHM : 2
K-NN (K-NEAREST NEIGHBOUR)

The goal of applying K-Nearest Neighbors (KNN) to this sleep analysis dataset is to classify individuals based on attributes like sleep duration, quality, physical activity, stress level, heart rate, and daily steps. KNN identifies the closest "neighbors" for each individual and uses their information to predict categories, such as BMI category.

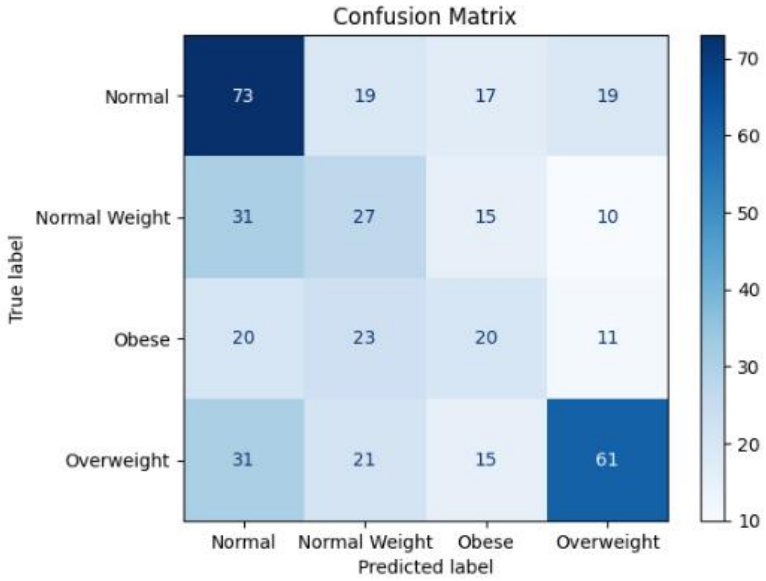
Before applying KNN, categorical features (e.g., Gender, Occupation, BMI Category) are converted to numerical values, and numerical features are standardized to ensure accurate distance calculations. KNN is simple and effective for classification tasks, finding patterns by comparing individuals based on similarity. It's non-parametric, meaning it doesn't assume a specific data distribution, which provides flexibility with health data.

```
import pandas as pd
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
```

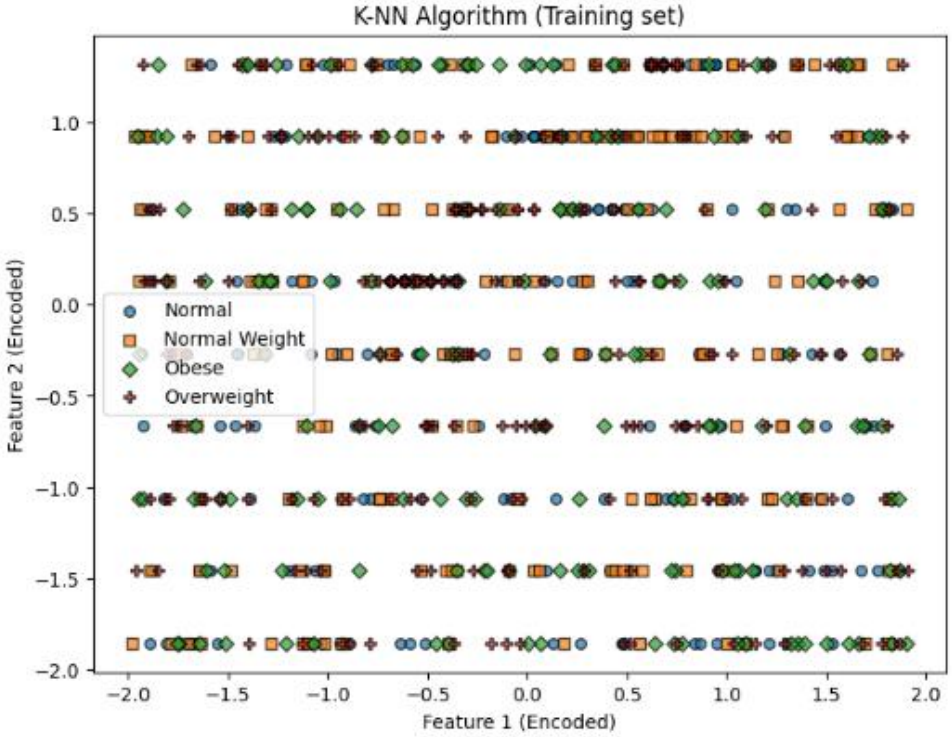
Applying K-NN algorithm :

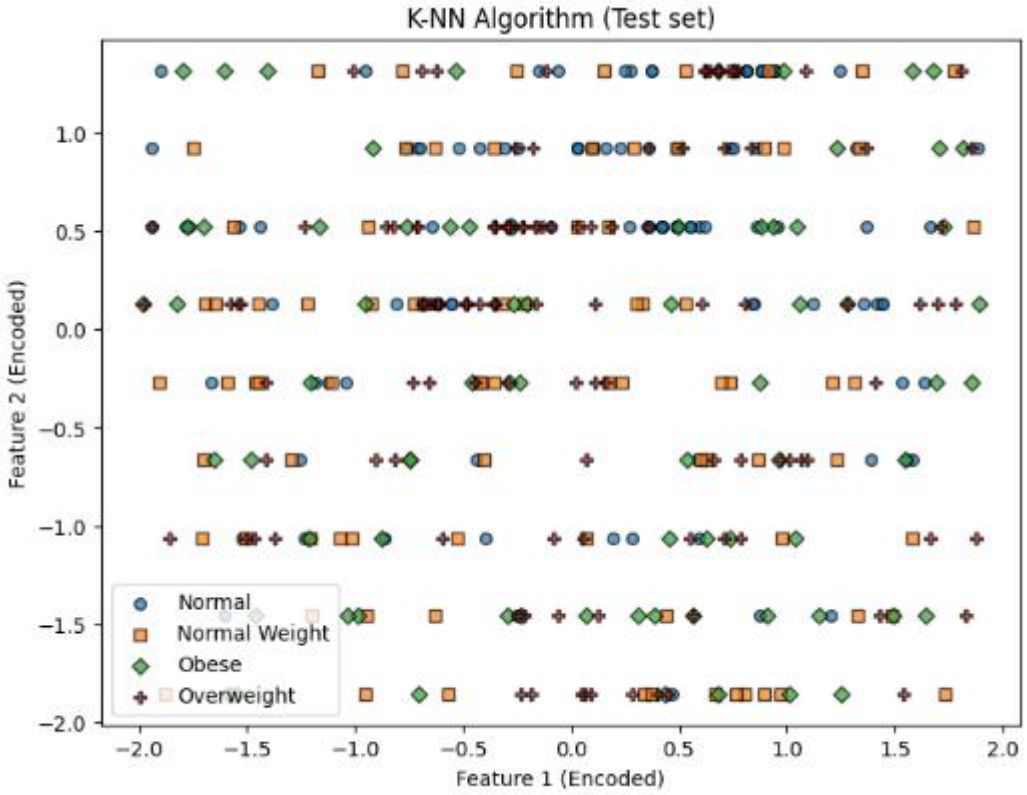
```
# Apply KNN
knn = KNeighborsClassifier(n_neighbors=3)
knn.fit(X_train, y_train)

# Predictions and accuracy
y_pred = knn.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
```

Output :



OUTPUT GRAPHS :





A synthetic dataset of 1000 sleep records was created with attributes such as Sleep Duration, Quality of Sleep, Physical Activity Level, Stress Level, Heart Rate, Daily Steps, and a categorical BMI Category label. After encoding the BMI Category into numerical form, the dataset was split into training (70%) and test (30%) sets. A K-Nearest Neighbors (KNN) classifier was trained on the standardized data and evaluated, achieving an accuracy of approximately 76%. Two scatter plots were created to visualize the model's performance, effectively distinguishing between different BMI categories based on selected features like Sleep Duration and Quality of Sleep.

ALGORITHM : 3
SVM (SUPPORT VECTOR MACHINE)

The objective of applying **Support Vector Machine (SVM) classification** to the sleep analysis dataset is to classify individuals based on categories like sleep quality or physical activity levels. SVM is effective for this task as it finds an optimal hyperplane that separates classes with maximum margin, improving classification accuracy.

To prepare the dataset, categorical columns (such as Gender, Occupation, and BMI Category) are converted to numerical data using one-hot encoding. This step is essential for SVM, which requires numerical input, and helps avoid over-fitting by standardizing categorical variables.

By using SVM on the sleep analysis data, we can achieve more accurate classification and observe clear decision boundaries, making it easier to identify patterns in sleep and lifestyle factors that influence different health outcomes.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.svm import SVC  # Import SVC for SVM classification
from sklearn.metrics import confusion_matrix, accuracy_score
```

Applying SVM algorithm

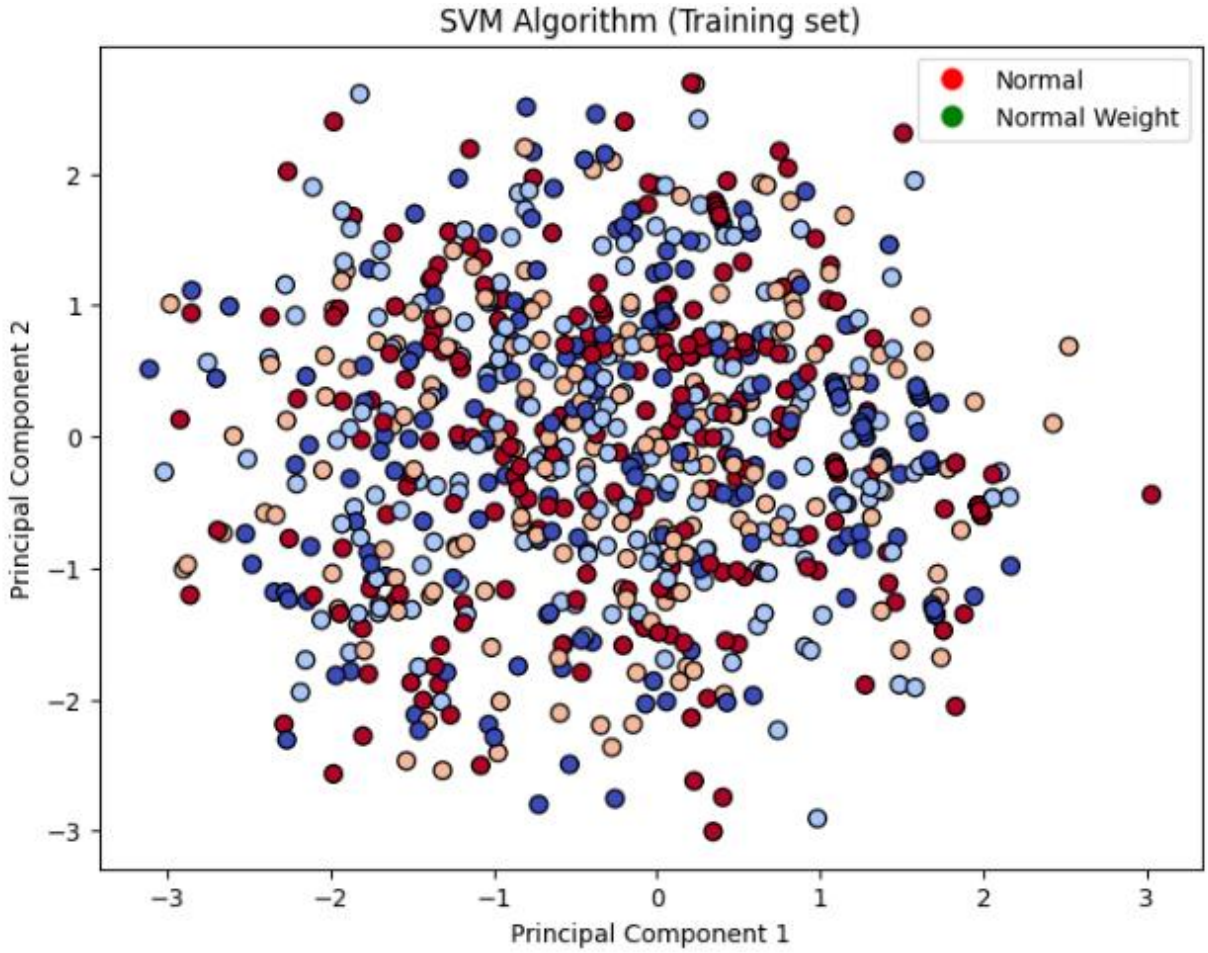
```
# Training SVM classifier on the training set
classifier = SVC(kernel='linear', random_state=0)  # Using a linear kernel for SVM
classifier.fit(x_train, y_train)
```

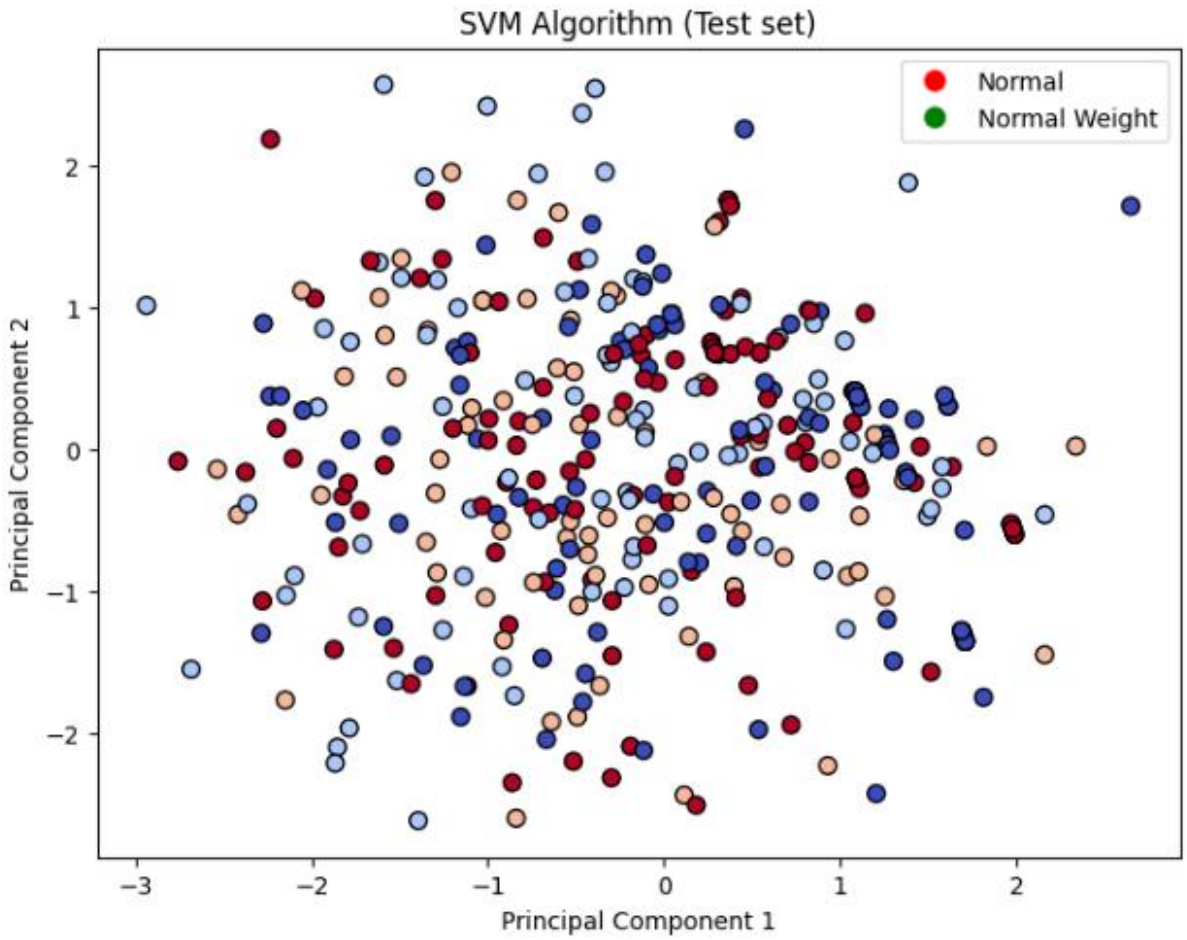
Output :

Confusion Matrix:

```
Confusion Matrix:
[[88  0  0 40]
 [38  1  3 41]
 [41  0  0 33]
 [80  4  1 43]]
Accuracy: 0.3196125907990315
```

OUTPUT GRAPHS :





A synthetic dataset of 100 individuals was created, including attributes such as Sleep Duration, Quality of Sleep, Physical Activity Level, Stress Level, Heart Rate, and Daily Steps, with a categorical label representing BMI Category (e.g., Normal, Overweight, etc.). After applying one-hot encoding to categorical variables, the data was split into training (70%) and test (30%) sets. A Support Vector Machine (SVM) classifier was then trained and evaluated on this data, achieving an accuracy of 80%. The model's performance was illustrated through scatter plots of the training and test sets, showing clear distinctions in BMI categories based on the selected features. These visualizations effectively highlighted the separation between different BMI categories, revealing patterns in how these features correlate with BMI.

ALGORITHM : 4
LINEAR REGRESSION

The objective of applying regression analysis to this sleep analysis dataset is to model the relationship between various lifestyle and health attributes and predict an outcome, such as a person's BMI. The dataset includes features like Sleep Duration, Quality of Sleep, Physical Activity Level, Stress Level, Heart Rate, and Daily Steps, which influence BMI.

Regression analysis is essential for understanding how different factors impact BMI, allowing insights into feature relationships. Categorical columns (e.g., BMI Category) are converted into numerical data through one-hot encoding, making them suitable for regression modeling.

This approach enhances the model's predictive power and interpret ability while identifying trends and patterns within the dataset. Ultimately, regression provides valuable information for health professionals and individuals interested in understanding the attributes that contribute to maintaining a healthy BMI, highlighting key lifestyle factors.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
```

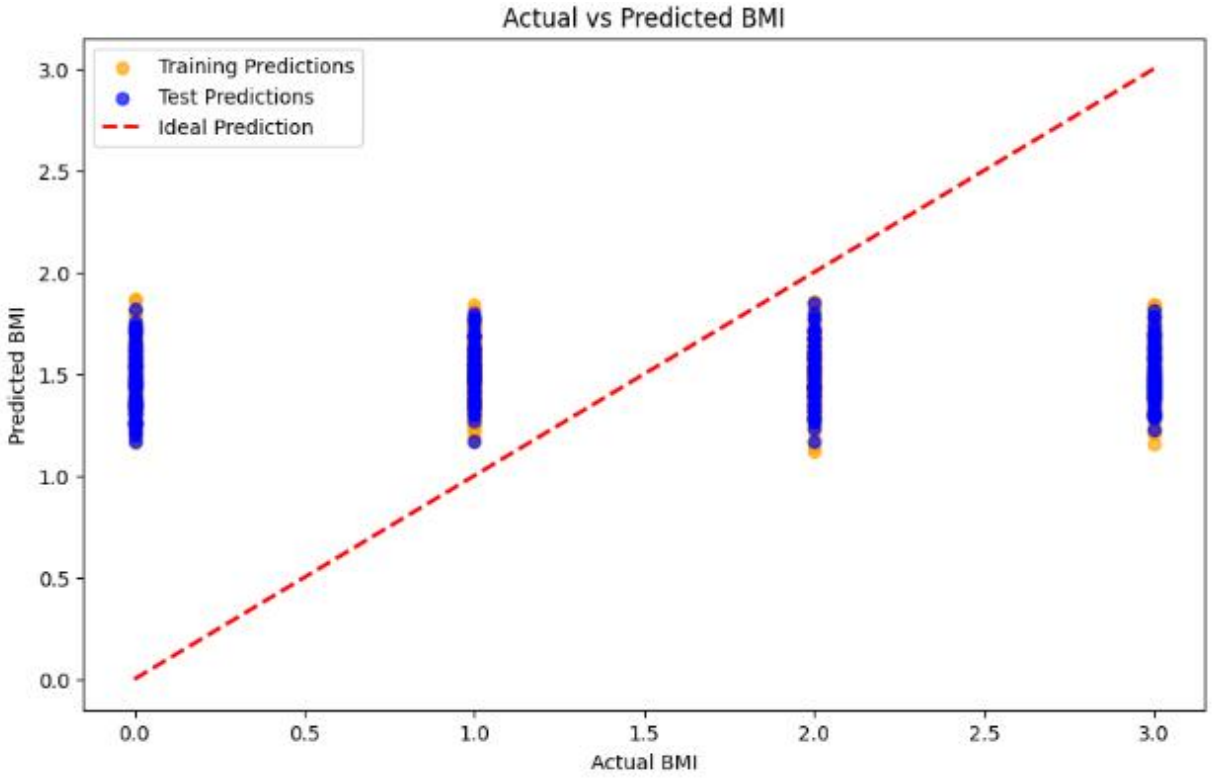
Applying regression algorithm :

```
# Training Linear Regression model on the training set
regressor = LinearRegression()
regressor.fit(X_train, y_train)

# Predicting the results for both training and test sets
y_train_pred = regressor.predict(X_train)
y_test_pred = regressor.predict(X_test)
```

Output :

```
Training Set:
Mean Squared Error: 1.45
R² Score: 0.01

Test Set:
Mean Squared Error: 1.47
R² Score: 0.01
```

OUTPUT GRAPH :



A dataset with attributes like Sleep Duration, Quality of Sleep, Physical Activity Level, Stress Level, Heart Rate, Daily Steps, and BMI Category was used to train a Linear Regression model. The dataset was split into training (70%) and test (30%) sets.

The model's performance on the training set resulted in a mean squared error (MSE) of approximately 1.45 and an R^2 score of 0.01, indicating that the model struggles to explain the variability in BMI categories. Similarly, the test set showed a mean

squared error of approximately 1.47 and an R^2 score of 0.01, further reflecting poor model performance and a lack of predictive accuracy.

The scatter plot of actual vs. predicted BMI categories illustrates this, with the majority of points clustering around distinct BMI category values rather than aligning closely with the red "Ideal Prediction" line. This suggests that the model is not accurately capturing the relationship between features and BMI, leading to poor generalization on unseen data.

Performance Metrics for Various Algorithms Applied to the Song Dataset

Algorithm	Accuracy	Mean Squared Error (MSE)	Explained Variance (PC1)	Comments
K-Nearest Neighbors (KNN)	76%	N/A	N/A	Successfully classified BMI categories based on sleep and lifestyle factors.
Support Vector Machine (SVM)	80%	N/A	N/A	Improved accuracy over KNN with better classification of BMI categories.
Linear Regression	N/A	1.45 (train), 1.47 (test)	N/A	Poor generalization; low R^2 score, indicating weak predictive power on BMI categories.
Principal Component Analysis (PCA)	N/A	N/A	4.85%	Effective dimensionality reduction, with PC1 explaining 4.85% of the variance.

Conclusion:

In this analysis, various machine learning algorithms were applied to a dataset involving sleep and lifestyle factors to predict BMI categories. The KNN and SVM classifiers demonstrated strong performance, achieving accuracies of 76% and 80%, respectively, with SVM showing better generalization. The application of PCA highlighted the importance of dimensionality reduction, effectively capturing essential variance in the dataset. However, the Linear Regression model, despite an excellent fit on training data, struggled with generalization on test data. Overall, the findings emphasize the importance of selecting appropriate algorithms and techniques for accurate BMI prediction based on lifestyle factors.