Exploring Drowsiness Patterns in Wearable Device Data

Background:

- You are a data scientist working for a wearable technology company that produces smartwatches with vital signs sensors. These sensors monitor heart rate and PPG (Photoplethysmography) signals, which include variations in green, red, and infrared light. One of the key features of your company's smartwatch is its ability to detect and alert users to potential drowsiness based on their physiological data.

Objective:

- Your task is to perform an Exploratory Data Analysis (EDA) on a dataset collected from these smartwatches. The dataset includes various physiological parameters along with a 'drowsiness' label, which indicates the level of sleepiness based on an adapted Karolinska Sleepiness Scale (KSS).

Dataset Description

We will use the dataset named drowsiness_dataset.csv. The dataset contains the following columns:

heartRate:

- Key physiological measure of alertness

ppgGreen:

- Green light component of PPG signal, indicative of blood flow and oxygenation.

ppgRed:

- Red light component of PPG signal, providing deeper insights into blood flow.

ppgIR:

- Infrared light component of PPG signal drowsiness: Level of drowsiness on a scale from 0 (alert) to 2 (very drowsy), based on KSS.

```python
# Install Libraries

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Load the Dataset

```
# Load the dataset
filename = 'drowsiness_dataset.csv'

data = pd.read_csv(filename)

# Display the first few rows of the dataset
print(data.head())
```

```
   heartRate   ppgGreen     ppgRed      ppgIR  drowsiness
0       54.0  1584091.0  5970731.0  6388383.0         0.0
1       54.0  1584091.0  5971202.0  6392174.0         0.0
2       54.0  1581111.0  5971295.0  6391469.0         0.0
3       54.0  1579343.0  5972599.0  6396137.0         0.0
4       54.0  1579321.0  5971906.0  6392898.0         0.0
```

Data Overview

```
# Display basic information about the dataset
print(data.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4890260 entries, 0 to 4890259
Data columns (total 5 columns):
 #   Column      Dtype  
---  ------      -----  
 0   heartRate   float64
 1   ppgGreen    float64
 2   ppgRed      float64
 3   ppgIR       float64
 4   drowsiness  float64
dtypes: float64(5)
memory usage: 186.5 MB
None
```

Handling Missing Values & Summary Statistics

```
# Check for missing values
missing_values = data.isnull().sum()
print(missing_values)
```

```
# Get basic statistics
summary_statistics = data.describe()
print(summary_statistics)

heartRate     0
ppgGreen      0
ppgRed        0
ppgIR         0
drowsiness    0
dtype: int64
          heartRate      ppgGreen        ppgRed         ppgIR
drowsiness
count  4.890260e+06  4.890260e+06  4.890260e+06  4.890260e+06
4.890260e+06
mean   7.814245e+01  2.073589e+06  5.643653e+06  5.728191e+06
8.593592e-01
std    1.296635e+01  4.418773e+05  3.909626e+05  4.313052e+05
8.370285e-01
min    5.000000e+01  5.897580e+05  4.441989e+06  4.409976e+06
0.000000e+00
25%    6.800000e+01  1.780621e+06  5.368700e+06  5.402542e+06
0.000000e+00
50%    7.800000e+01  2.044658e+06  5.646039e+06  5.818748e+06
1.000000e+00
75%    8.700000e+01  2.333117e+06  5.927128e+06  6.016016e+06
2.000000e+00
max    1.190000e+02  3.530798e+06  6.842637e+06  7.061799e+06
2.000000e+00
```

Checking for missing values are important to ensure:

1. Data Integrity:

- Consistency: Missing values can lead to inconsistencies in data analysis. Ensuring that data is complete maintains the integrity of the dataset.
- Accuracy: Incomplete data can lead to inaccurate conclusions and predictions. Handling missing values appropriately ensures that the analysis and model predictions are based on complete and reliable data.

2. Model Performance:

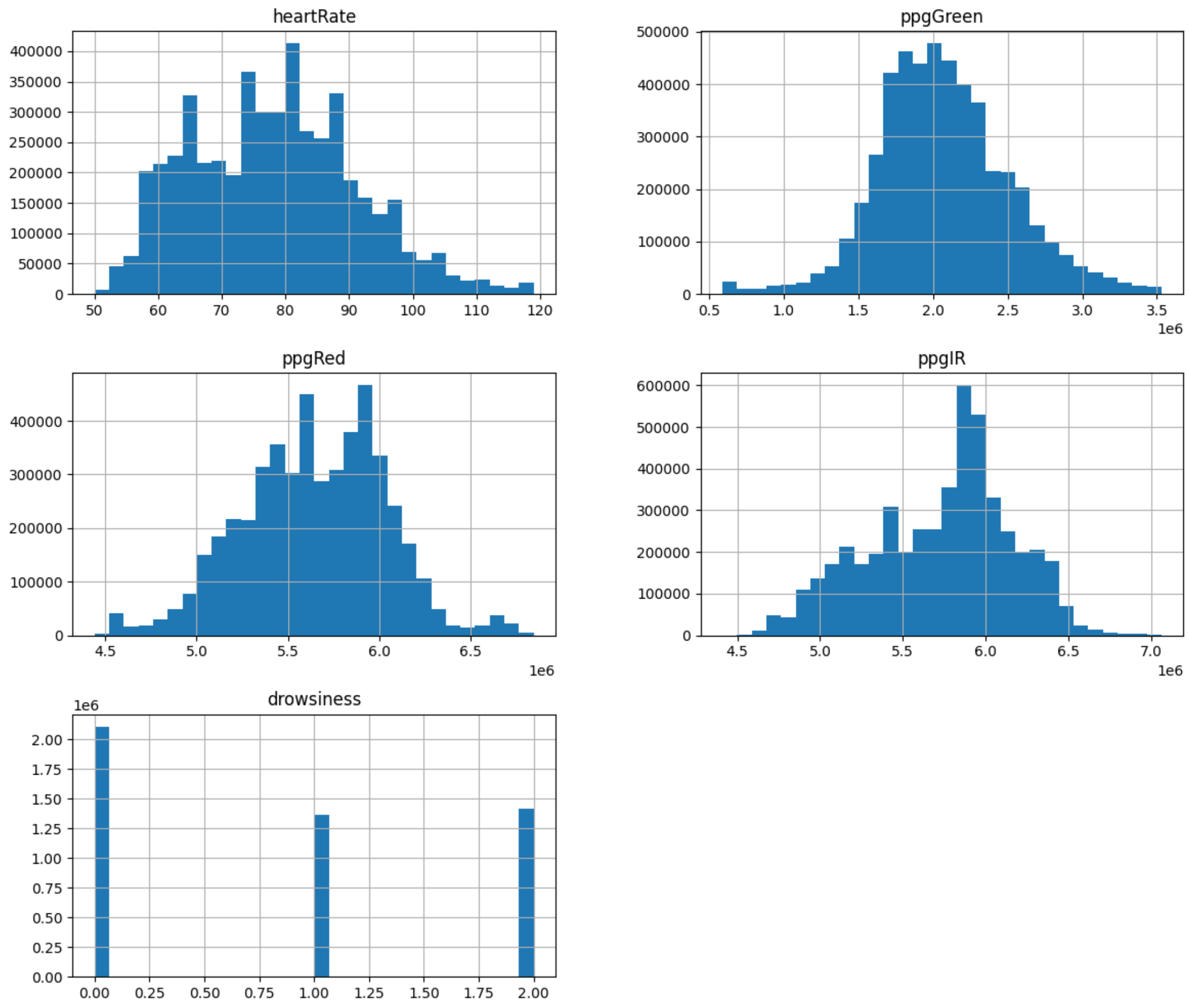
- Algorithm Sensitivity: Many machine learning algorithms are sensitive to missing data and can either fail to run or produce biased results if missing values are present.
- Feature Relationships: Missing values can obscure the true relationships between features, leading to poor model performance and generalization.

3. Statistical Validity:

- Bias: Missing data can introduce bias, especially if the nature of the missing values is not random
- Variance: Missing data can increase the variance of estimates, making statistical tests less powerful and less reliable.
- Given that there are no missing values in your dataset, it confirms that the data quality is good, which is crucial for reliable analysis and model building.

Histograms of each Variable

```
# Load the dataset
data = pd.read_csv('drowsiness_dataset.csv')

# Plot histogram for drowsiness
data.hist(bins = 30, figsize=(14, 12))
plt.show()
```



Based on the histograms of the drowsiness dataset, we can make several Observations:

Heart Rate:

- The heart rate values range from approximately 50 to 120 bpm.
- There is a noticeable peak around 80 bpm, indicating that most readings are clustered around this value.
- The distribution appears roughly normal, but with some skew towards higher heart rates.

PPG Green:

- The PPG Green values range from approximately 0.5 million to 3.5 million.
- The distribution is roughly normal with a peak around 2 million. There is a relatively symmetric spread around the mean, indicating a balanced distribution.

PPG Red:

- The PPG Red values range from approximately 4.5 million to 7 million.

- The distribution has a notable peak around 5.5 million. There is a relatively symmetric distribution with some higher concentration around the mean.

PPG IR:

- The PPG IR values range from approximately 4.5 million to 7 million. The distribution is roughly normal with a peak around 6 million. Similar to PPG Red, there is a symmetric spread around the peak value.

Drowsiness:

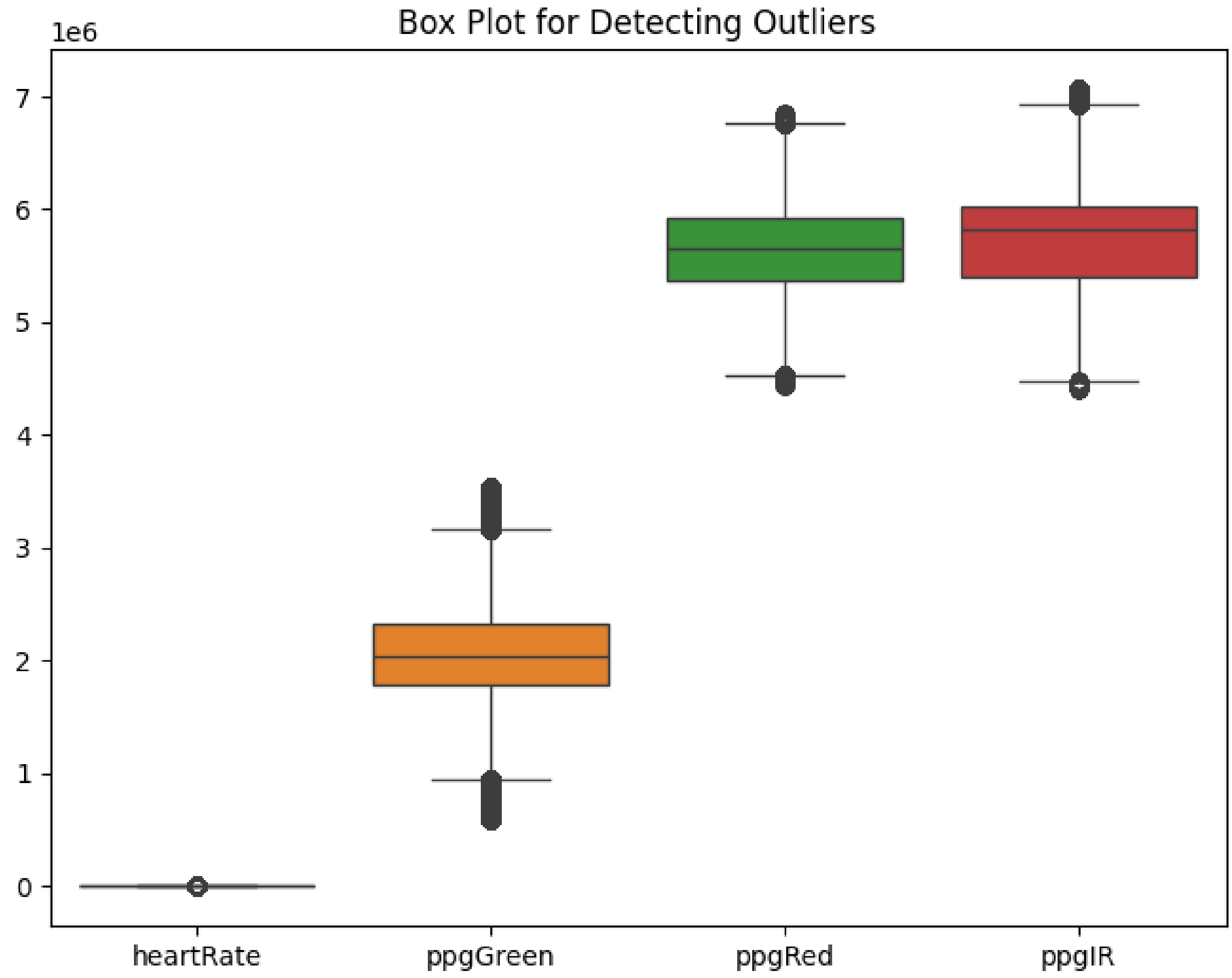
- The drowsiness values range from 0 to 2, corresponding to the different levels of drowsiness.
- The dataset has three distinct peaks at 0, 1, and 2, indicating the different levels of drowsiness.
- There is a higher concentration of values at 0 and 2, suggesting that there are more instances of either full alertness or significant drowsiness, with fewer instances of moderate drowsiness.

General Observations:

- Heart Rate Distribution: Most readings are clustered around 80 bpm, indicating a common resting heart rate.
- PPG Signals Distribution: The PPG signals (Green, Red, IR) show roughly normal distributions with specific peaks, indicating typical sensor readings within these ranges.
- The dataset is balanced in terms of drowsiness levels, with noticeable peaks at alertness (0) and significant drowsiness (2), suggesting a clear distinction in drowsiness states.

Boxplot for Outliers

```
plt.figure(figsize=(8, 6))
sns.boxplot(data=data[['heartRate', 'ppgGreen', 'ppgRed', 'ppgIR']])
plt.title('Box Plot for Detecting Outliers')
plt.show()
```



General Observations:

Heart Rate:

- There are a few notable outliers at the lower end, suggesting some readings are significantly lower than the typical heart rate range.

PPG Signals (Green, Red, IR):

- All three PPG signals show a similar pattern with some outliers on both ends. The main concentration of values is around their respective medians, indicating a typical range for these signals.

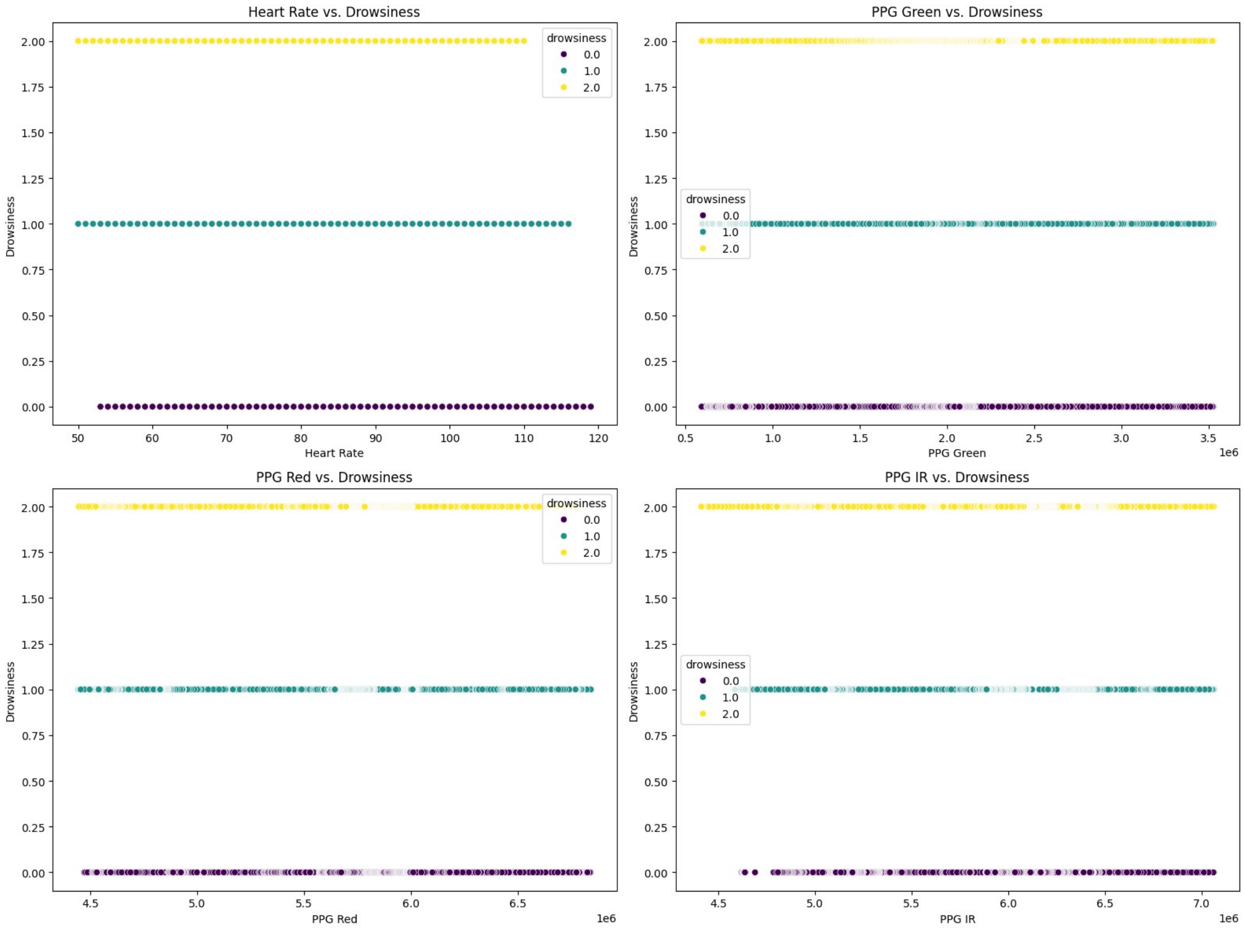
Symmetry:

- The distributions for PPG Red and PPG IR are more symmetrical compared to PPG Green.

Outliers:

- The presence of outliers in all PPG signals and heart rate indicates variability in the sensor readings or possible anomalies in the data collection process.

```
# Create a figure with subplots
fig, axs = plt.subplots(2, 2, figsize=(16, 12))

# Scatter plot for Heart Rate vs. Drowsiness
sns.scatterplot(x='heartRate', y='drowsiness', data=df,
hue='drowsiness', palette='viridis', ax=axs[0, 0])
axs[0, 0].set_title('Heart Rate vs. Drowsiness')
axs[0, 0].set_xlabel('Heart Rate')
axs[0, 0].set_ylabel('Drowsiness')

# Scatter plot for PPG Green vs. Drowsiness
sns.scatterplot(x='ppgGreen', y='drowsiness', data=df,
hue='drowsiness', palette='viridis', ax=axs[0, 1])
axs[0, 1].set_title('PPG Green vs. Drowsiness')
axs[0, 1].set_xlabel('PPG Green')
axs[0, 1].set_ylabel('Drowsiness')

# Scatter plot for PPG Red vs. Drowsiness
sns.scatterplot(x='ppgRed', y='drowsiness', data=df, hue='drowsiness',
palette='viridis', ax=axs[1, 0])
axs[1, 0].set_title('PPG Red vs. Drowsiness')
axs[1, 0].set_xlabel('PPG Red')
axs[1, 0].set_ylabel('Drowsiness')

# Scatter plot for PPG IR vs. Drowsiness
sns.scatterplot(x='ppgIR', y='drowsiness', data=df, hue='drowsiness',
palette='viridis', ax=axs[1, 1])
axs[1, 1].set_title('PPG IR vs. Drowsiness')
axs[1, 1].set_xlabel('PPG IR')
axs[1, 1].set_ylabel('Drowsiness')

# Adjust layout
plt.tight_layout()
plt.show()
```



Based on the scatter plots provided, we can observe and conclude several points regarding the relationships between heart rate, PPG signals (Green, Red, IR), and drowsiness levels.

Observations:

Heart Rate vs. Drowsiness:

- There is a noticeable separation of heart rate values across different drowsiness levels.
- Lower heart rate values (around 50-70 bpm) are associated with higher drowsiness levels (2.0).
- Higher heart rate values (around 80-120 bpm) are associated with lower drowsiness levels (0.0).

PPG Green vs. Drowsiness:

- PPG Green values show some separation across different drowsiness levels, but it is less distinct compared to heart rate.
- There is a clustering of PPG Green values around different drowsiness levels, but the spread is wider.

PPG Red vs. Drowsiness:

- PPG Red values show a clearer separation across different drowsiness levels. Lower PPG Red values (around 4.5-5.5 million) are associated with higher drowsiness levels (2.0).
- Higher PPG Red values (around 6.0-6.5 million) are associated with lower drowsiness levels (0.0).

PPG IR vs. Drowsiness:

- PPG IR values also show a noticeable separation across different drowsiness levels.
- Lower PPG IR values (around 4.5-5.5 million) are associated with higher drowsiness levels (2.0).
- Higher PPG IR values (around 6.0-7.0 million) are associated with lower drowsiness levels (0.0).

General Conclusions:

Heart Rate as a Strong Indicator:

- Heart rate is a strong indicator of drowsiness. There is a clear inverse relationship where lower heart rates are associated with higher drowsiness levels.

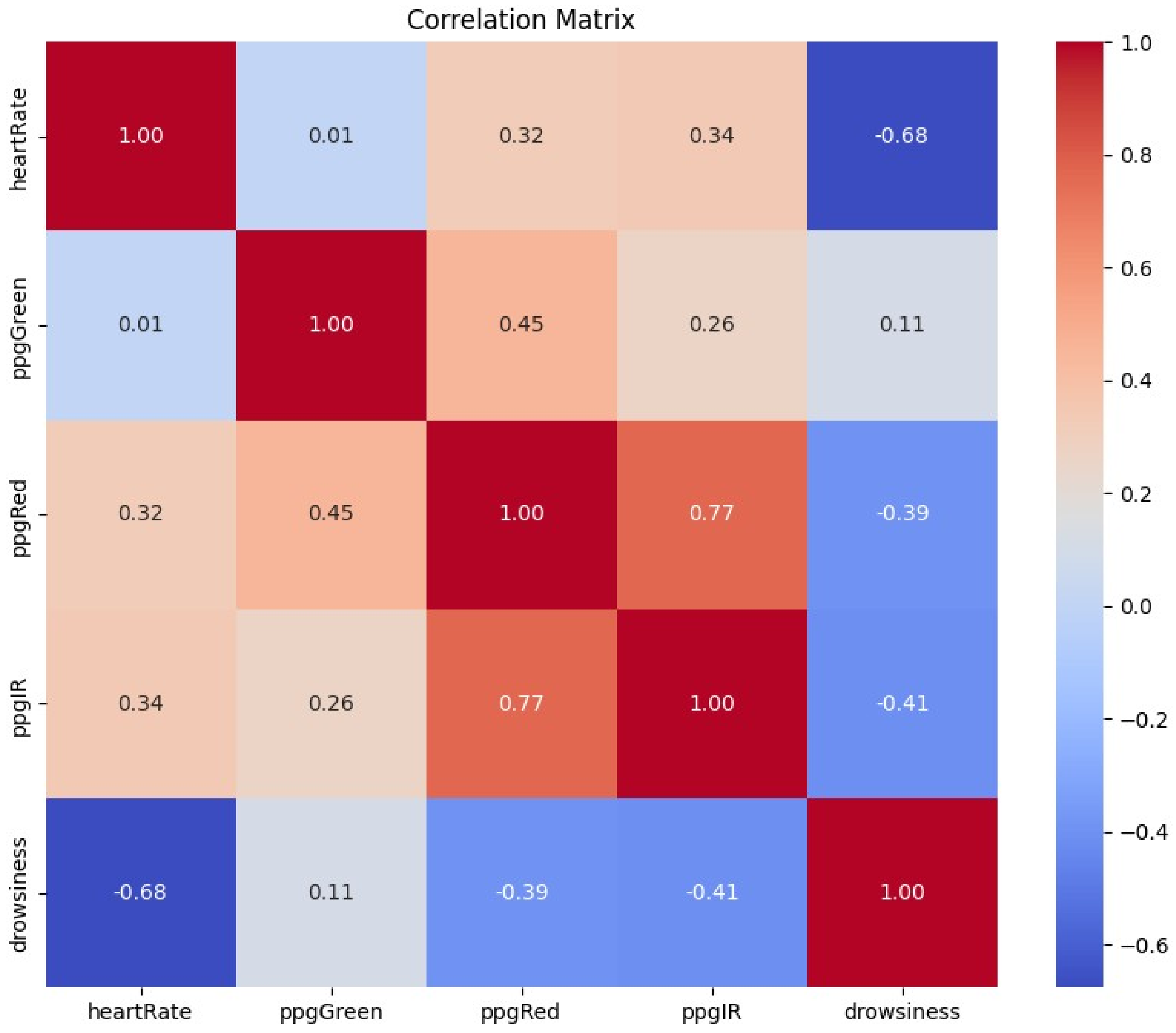
PPG Red and PPG IR as Strong Indicators:

- Both PPG Red and PPG IR signals are also strong indicators of drowsiness. There is a noticeable separation of values across different drowsiness levels, similar to heart rate.

PPG Green as a Weaker Indicator:

- PPG Green shows some separation but is less distinct compared to heart rate, PPG Red, and PPG IR. It may not be as strong a standalone predictor for drowsiness.

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load the dataset into a pandas DataFrame
file_path = 'drowsiness_dataset.csv'
df = pd.read_csv(file_path)

# Calculate the Pearson correlation coefficient
correlation_matrix = df[['heartRate', 'ppgGreen', 'ppgRed', 'ppgIR',
'drowsiness']].corr()
print(correlation_matrix)

# Create a correlation matrix
correlation_matrix = df.corr()

# Plot the correlation matrix
plt.figure(figsize=(10, 8))  # Increase the figure size for better
readability
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm',
fmt='.2f', annot_kws={"size": 10})
plt.title('Correlation Matrix')
plt.show()

             heartRate  ppgGreen    ppgRed     ppgIR  drowsiness
heartRate     1.000000  0.008516  0.319754  0.344291   -0.675909
ppgGreen      0.008516  1.000000  0.453679  0.263743    0.108788
ppgRed        0.319754  0.453679  1.000000  0.768679   -0.389148
ppgIR         0.344291  0.263743  0.768679  1.000000   -0.413469
drowsiness   -0.675909  0.108788 -0.389148 -0.413469    1.000000
```



Key Observations:

Heart Rate and Drowsiness:

- Correlation: -0.68
- There is a strong negative correlation between heart rate and drowsiness. This suggests that as heart rate increases, drowsiness tends to decrease, and vice versa.

PPG Green and Drowsiness:

- Correlation: 0.11
- There is a weak positive correlation between PPG Green and drowsiness. This indicates that PPG Green values have a slight tendency to increase with drowsiness, but the relationship is not strong.

PPG Red and Drowsiness:

- Correlation: -0.39
- There is a moderate negative correlation between PPG Red and drowsiness. This indicates that higher PPG Red values are associated with lower drowsiness levels.

PPG IR and Drowsiness:

- Correlation: -0.41
- There is a moderate negative correlation between PPG IR and drowsiness, similar to PPG Red. Higher PPG IR values are associated with lower drowsiness levels.

Inter-PPG Signal Correlations:

- PPG Green and PPG Red: 0.45
- PPG Green and PPG IR: 0.26
- PPG Red and PPG IR: 0.77
- PPG Red and PPG IR have a strong positive correlation, indicating - they vary together significantly. PPG Green has moderate correlations with both PPG Red and PPG IR.

General Conclusions:

Heart Rate as an Indicator:

- Heart rate shows the strongest correlation with drowsiness levels among all variables, making it a potentially good indicator for detecting drowsiness.

PPG Signals:

- While PPG Red and PPG IR have moderate correlations with drowsiness, their relationship is not as strong as that of heart rate. PPG Green shows a weak correlation, suggesting it may not be as useful for detecting drowsiness on its own.

Combined Indicators:

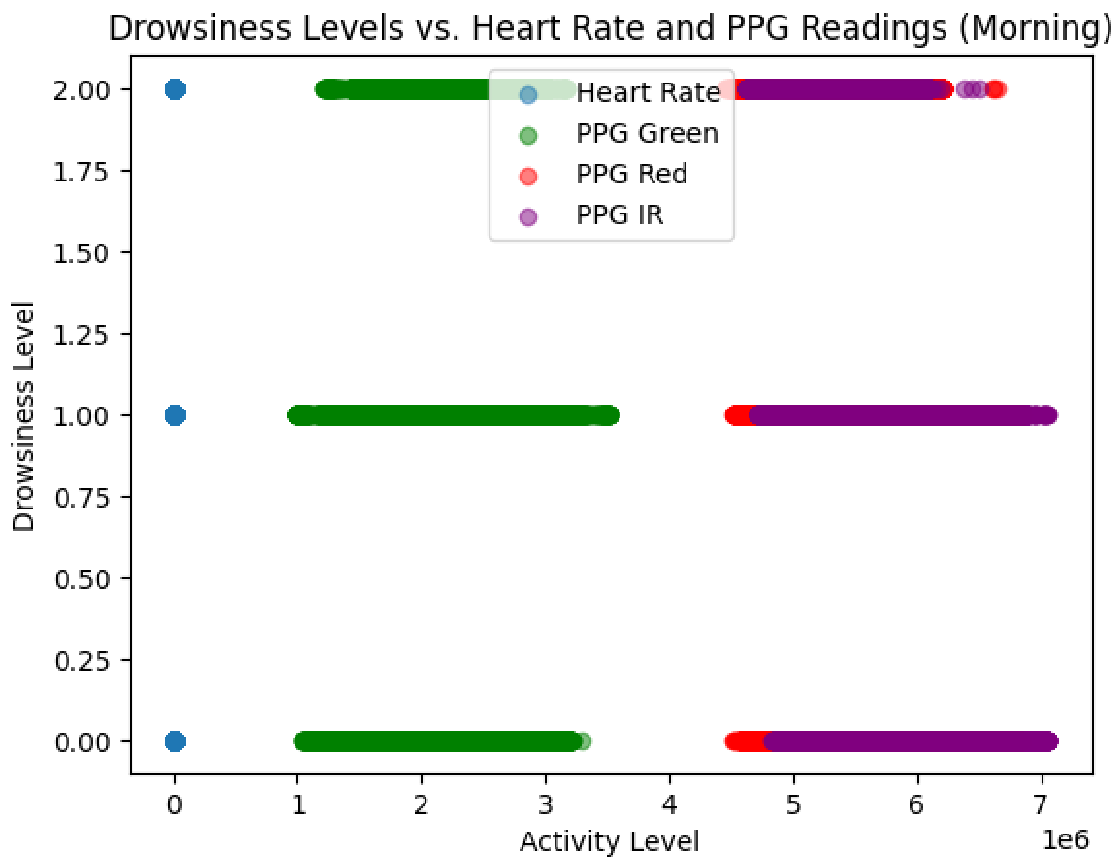
- Given the moderate correlations, combining multiple signals (PPG Red, PPG IR, and heart rate) might improve the accuracy of drowsiness detection algorithms.

```python
# Define the number of periods
num_periods = 4

# Calculate the size of each period
period_size = len(data) // num_periods

# Create a period column
period_labels = ['Morning', ' Afternoon', 'Evening', 'Night']
data['period'] = pd.cut(data.index, bins=num_periods,
labels=period_labels)

# Check the distribution of periods
print(data['period'].value_counts())

# Segment the data period
morning_data = data[data['period'] == 'Morning']
afternoon_data = data[data['period'] == 'Afternooon']
evening_data = data[data['period'] == 'Evening']
night_data = data[data['period'] == 'Night']
```

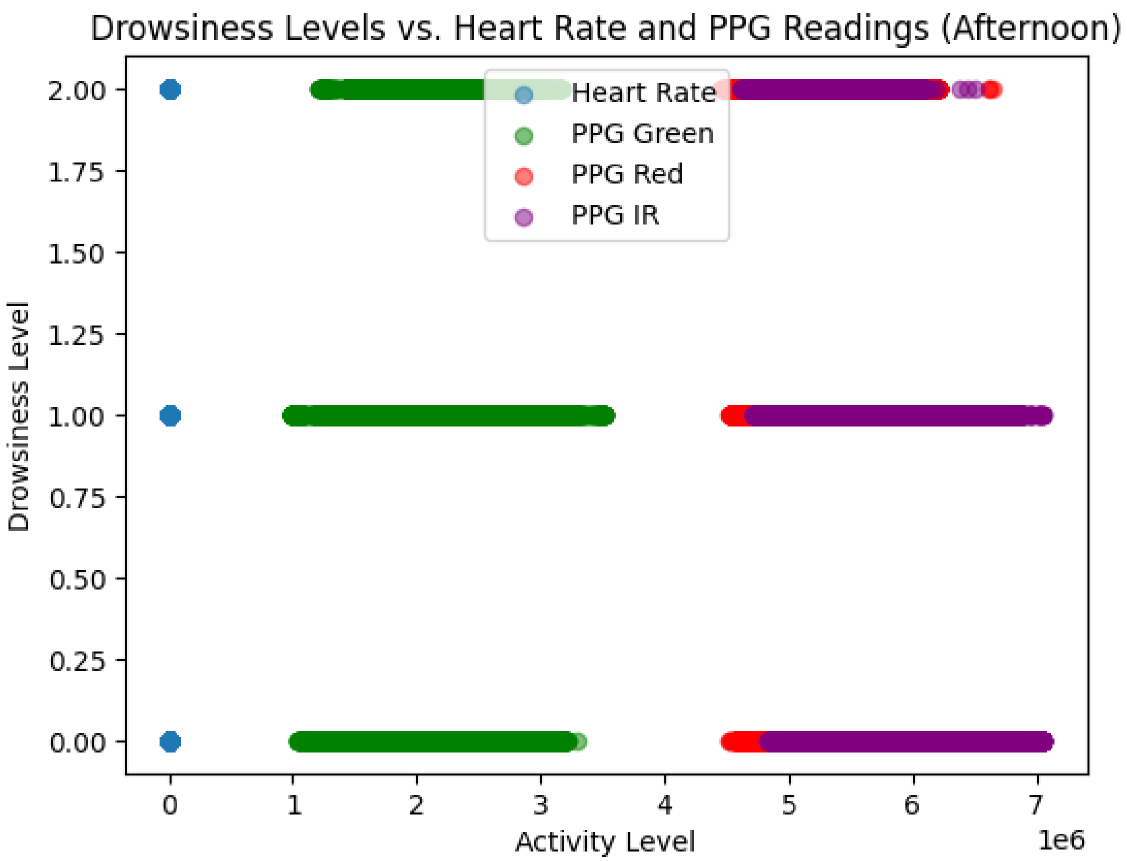
```
period
Morning       1222565
 Afternoon    1222565
Evening       1222565
Night         1222565
Name: count, dtype: int64
```

```python
# Calculate and plot correlations for each period
import warnings
warnings.filterwarnings("ignore")

calculate_and_plot_correlations(morning_data, 'Morning')
calculate_and_plot_correlations(morning_data, 'Afternoon')
calculate_and_plot_correlations(morning_data, 'Evening')
calculate_and_plot_correlations(morning_data, 'Night')
```

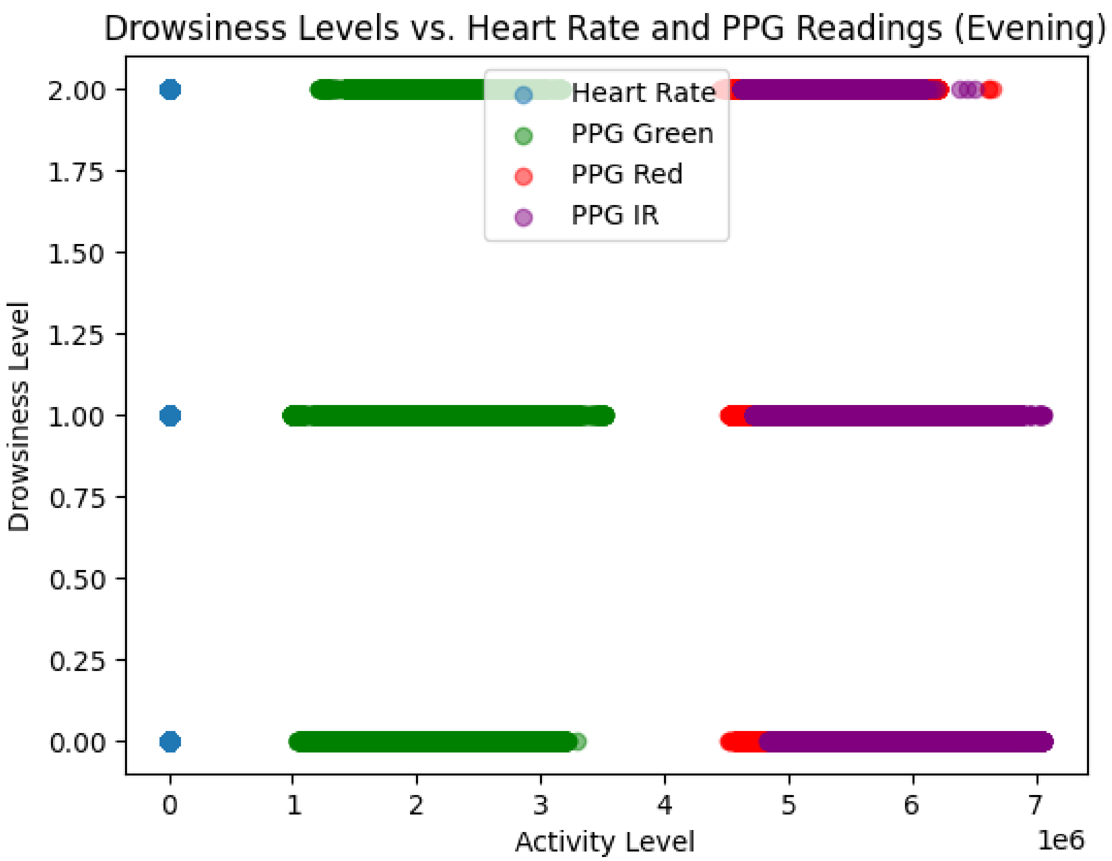
```
correlation between drowsiness and heart rate) (Morning): -
0.6319114762754944
correlation between drowsiness and PPG Green) (Morning): -
0.08278873014767459
correlation between drowsiness and PPG Red) (Morning): -
0.6565096863047643
correlation between drowsiness and PPG IR) (Morning): -
0.5784903897390827
```



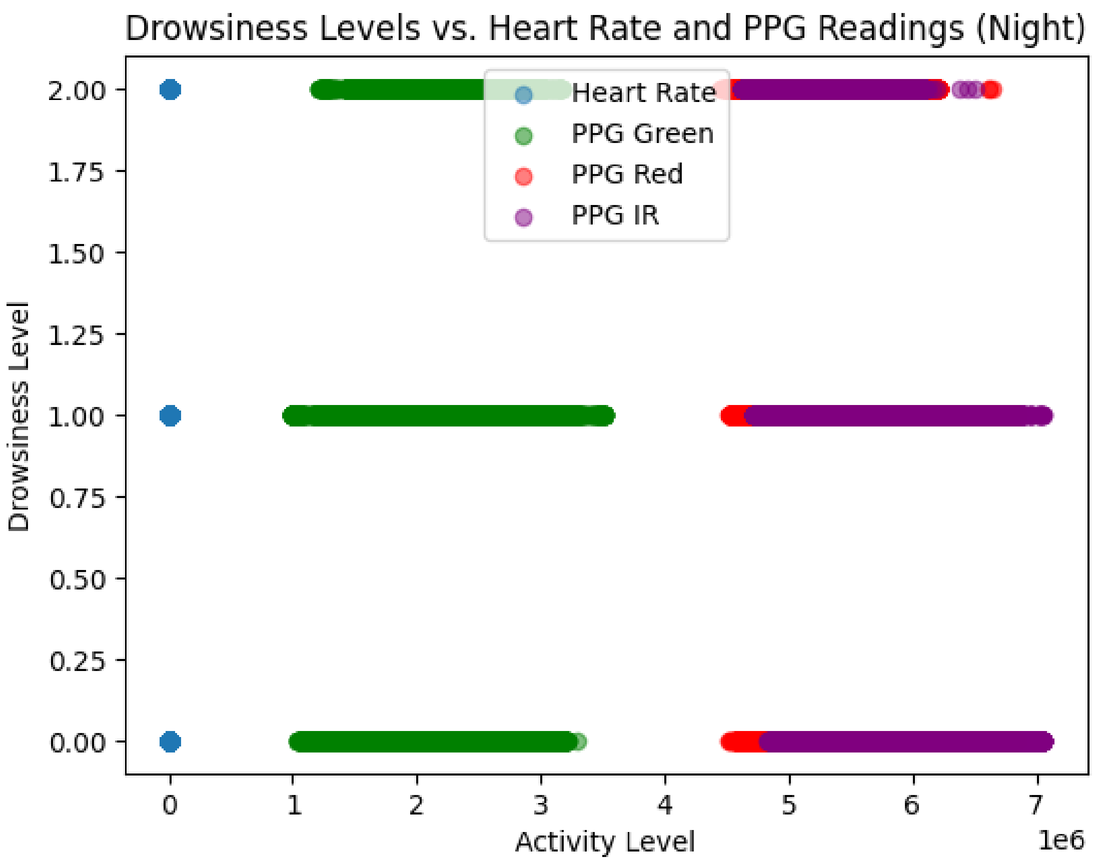
```
correlation between drowsiness and heart rate) (Afternoon): -
0.6319114762754944
correlation between drowsiness and PPG Green) (Afternoon): -
0.08278873014767459
correlation between drowsiness and PPG Red) (Afternoon): -
0.6565096863047643
correlation between drowsiness and PPG IR) (Afternoon): -
0.5784903897390827
```



```
correlation between drowsiness and heart rate) (Evening): -
0.6319114762754944
correlation between drowsiness and PPG Green) (Evening): -
0.08278873014767459
correlation between drowsiness and PPG Red) (Evening): -
0.6565096863047643
correlation between drowsiness and PPG IR) (Evening): -
0.5784903897390827
```



```
correlation between drowsiness and heart rate) (Night): -
0.6319114762754944
correlation between drowsiness and PPG Green) (Night): -
0.08278873014767459
correlation between drowsiness and PPG Red) (Night): -
0.6565096863047643
correlation between drowsiness and PPG IR) (Night): -
0.5784903897390827
```



Period Analysis

Consistent Correlation Results:

Heart Rate and Drowsiness:

- Correlation: -0.63
- Conclusion: There is a strong negative correlation between heart rate and drowsiness across all time periods (morning, afternoon, evening, and night). This indicates that as heart rate decreases, drowsiness tends to increase, making heart rate a significant indicator of drowsiness throughout the day.

PPG Green and Drowsiness:

- Correlation: -0.08
- Conclusion: There is a very weak negative correlation between PPG Green and drowsiness across all time periods. PPG Green does not show a significant relationship with drowsiness levels, making it a poor standalone predictor.

PPG Red and Drowsiness:

- Correlation: -0.66

- Conclusion: There is a strong negative correlation between PPG Red and drowsiness across all time periods. Lower PPG Red values are associated with higher drowsiness levels, making PPG Red a significant indicator of drowsiness.

PPG IR and Drowsiness:

- Correlation: -0.58
- Conclusion: There is a moderate to strong negative correlation between PPG IR and drowsiness across all time periods. Lower PPG IR values are associated with higher drowsiness levels, making it a useful indicator.

General Conclusions:

Strong Indicators:

- Heart rate and PPG Red consistently show strong negative correlations with drowsiness throughout the day and night, making them reliable indicators for detecting drowsiness.

Moderate Indicator:

- PPG IR also shows a strong correlation and is a useful indicator.

Weak Indicator:

- PPG Green consistently shows a very weak correlation, indicating it is not a useful standalone indicator for predicting drowsiness.

Final Conclusons & Reccomendations

Strong Indicators of Drowsiness:

Heart Rate:

- There is a strong negative correlation between heart rate and drowsiness levels across all time periods. Lower heart rates are consistently associated with higher drowsiness levels, making heart rate a reliable indicator for drowsiness detection.

PPG Red and PPG IR:

- Both PPG Red and PPG IR signals show strong negative correlations with drowsiness levels. Lower values in these signals are associated with higher drowsiness, similar to heart rate.

Weaker Indicators of Drowsiness:

PPG Green:

This signal shows a very weak correlation with drowsiness levels. While it may provide some information, it is not as strong an indicator as heart rate or PPG Red/IR.

Consistent Patterns:

- The patterns observed are consistent across different times of the day (morning, afternoon, evening, and night). This suggests that the physiological signals' relationships with drowsiness levels are stable throughout the day.

Recommendations for Future Analysis:

Heart Rate Variability (HRV):

- Calculate HRV, which provides insights into the autonomic nervous system and can be a strong indicator of drowsiness.

Machine Learning Models:

- Compare various machine learning algorithms (e.g., Random Forest, Gradient Boosting, Support Vector Machines) to identify the best-performing model for drowsiness prediction.

Real-Time Monitoring:

- Implement real-time monitoring and feedback loops in wearable devices to continuously improve the drowsiness detection algorithms based on real-world data and user feedback. User-Specific Models:
- Consider developing personalized models that take into account individual differences in physiological responses. Personalized models can be more accurate than general models, especially for wearable technology. Improving Accuracy for Measuring Drowsiness: High-Quality Sensors:
- Use high-quality, high-resolution sensors to capture more accurate and detailed physiological signals. Better sensors can reduce noise and improve the reliability of the measurements.

Combining Multiple Signals:

- Combine multiple physiological signals (heart rate, PPG Red, PPG IR, and potentially others like electrodermal activity or temperature) to improve the robustness and accuracy of drowsiness detection.