Sleep Quality Production

Bruce Metoyer
Computer Science
Data Science
metoyerb0612@students.bowiestate.edu

Isaiah Atcherson
Computer Science
Data Science
atchersoni0809@students.bowiestate.edu

Muna Chongo
Computer Science
Data Science
chongom0911@students.bowiestate.edu

***Abstract*—Sleep is essential to human health and well-being, yet approximately 30% of adults worldwide suffer from sleep deprivation, particularly among young adults and teenagers. Poor sleep quality has been linked to adverse mental, cognitive, and physiological outcomes, including increased risks of heart disease and diabetes. This project examines the relationship between sleep quality and physiological signals collected through wearable devices, including heart rate (HR), inter-beat interval (IBI), blood volume pulse (BVP), and other biometric indicators. Using machine learning techniques, we developed predictive models to analyze sleep quality metrics and identify correlations with sleep quality.**

Keywords—Sleep deprivation, Cognitive, Disease and diabetes, Heart Rate, Inter-beat interval, Blood Volume Pulse, Machine Learning, Correlations, Predictive Models

I. Introduction

Everyone in the world needs sleep. Good quality of sleep measures an individual's productivity, energy, and success. Poor sleep quality can negatively impact mental and cognitive functions while increasing the risks of heart disease and diabetes. Approximately 30% of adults worldwide don't get enough sleep, meaning that about one-third of the population is sleep-deprived, mainly young adults and teenagers. This can be attributed to schoolwork and college, working long shifts and multiple jobs, and family responsibilities. The prevalence of sleep deprivation is a pressing issue that requires immediate attention. Overall, health is vital, so taking care of your body by eating the right foods and getting the right amount of sleep can have a significant impact on you.

Several factors can contribute to both positive and negative sleep quality. This project will analyze multiple wearable device metrics, including an individual's heart rate, oxygen levels, movement, and ambient conditions, during sleep. Using machine learning (ML) techniques, we aim to develop a predictive model that identifies patterns in the given metrics and provides insights into their impact on sleep. Ultimately, the goal is to support the improvement of sleep health through data-driven analysis and personalized feedback.

II. Previous Studies

We obtained data that was originally synthetic and featured about 300+ rows of sleep data that was made up, so originally it was going to be used in the prediction analysis, but the next study was soon found that featured real data that can be included.

The next paper we chose has real data and examines the effects of wearable devices on sleep quality in healthy individuals. The study analyzes various sleep parameters using polysomnography and wearables to obtain the reliability of these devices in tracking sleep metrics. The goal of the paper is to assess the effectiveness of wearable devices in monitoring sleep quality against traditional methods, such as video-assisted polysomnography, in individuals ages 24 to 86. The article discusses the various steps done to record and track data, a topic we go over often in our class. The paper also examines the participants' bodily information, such as their medical history, preexisting sleep disorders, and BMI. Along with the mean of their sleep apnea appearances, and the amount of time they woke up.

III. Methodology

Isaiah Atcherson: Analysis of a person's sleep quality was done and an attempt to create a prediction model that predicted if a person was going to develop sleep disorders or not, depending on their preexisting data and apnea indexes was made, all of the metrics were recorded except for medical history. Examined was all of the participants who all have varying ages, gender, and sleep disorders. Originally, I attempted to have all of the sleep disorders be predicted on by one and observe if a user can get one based on each separate one a participant has. Unfortunately, the decision tree was extremely complicated and the metrics it showcased were difficult to understand. Also the testing accuracy of the prediction model was significantly lower than the training accuracy PCA was also used in order to determine if the low testing accuracy could be a problem with dimensionality and it provided a slight advance in the accuracy which will be demonstrated later in the results.

Bruce Metoyer: We analyzed a person's sleep quality based on wearable heart rate metrics. The data we examined included raw BVP signals, HR, and IBI metrics to extract patterns in sleep quality and evaluate their relationship with the medical history and sleep disorders of each patient. In total, we examined 20 different patients, all of whom varied in age, gender, medical history, and sleep disorders. We initially examined only five patients due to space constraints on my device; however, the experiment's results allowed me to free up more space and increase the number of patients to 20.

Muna Chongo: To identify patterns associated with low sleep quality, we performed demographic visualizations using the computed Sleep Quality Index (SQI). Participants were grouped by categorical features including gender, obstructive sleep apnea (OSA) status, age, and body mass index (BMI). We employed violin plots to observe the distribution of SQI within each group, and bar plots with standard deviation error bars to summarize mean SQI across categories. These visualizations were generated using the seaborn and matplotlib Python libraries. The purpose was to explore how demographic factors influence sleep quality and to support the development of more accurate predictive models.

A. Data Description

The data we used was called DREAMT, which is a dataset for Real-time sleep stage Estimation using Multisensory wearable technology. One hundred unique participants, ranging in age, were recruited from the Duke University Health System (DUHS) Sleep Disorder Laboratory. This study took place from May 2022 to September 2022. The primary purpose of the experiment is to detect and monitor patients' apnea events during sleep. These patients abstain from caffeine after noon of the day of PSG recording to ensure standard protocol. No caffeine intake monitoring on the participant's arrival in the sleep lab. The participants arrived at 8 PM on the scheduled day of their appointment and stayed in a hotel room for the night at the sleep lab. Four participants were accommodated per night, due to limited equipment. After consent was obtained upon arrival, the Empatica E4 wristband was placed on the participant's left wrist. Overnight PSG was performed using the Nihon Khoden Polysmith 1004 Data Management System (DMS), where data collection began at 11 PM and continued until 6 AM throughout the night. The E4 device was deactivated at the time of awakening. The data is 500 rows by 14 columns - Sleep Duration and Sleep Data.

E4 device collected:

1. Blood Volume Pulse **(BVP)** derived from the photoplethysmography **(PPG)** sensor
2. Accelerometry (**ACC)** in 3 axes,
    1. **ACC_X**
    2. **ACC_Y**
    3. **ACC_Z**
3. Electrodermal Activity **(EDA)**
4. Skin Temperature **(TEMP)**
5. Heart Rate **(HR)**: estimated from raw **BVP**
6. Inter-beat Interval **(IBI)**: estimated from raw BVP
7. Sleep-stage labels, derived from **PSG**, are recorded every 30 seconds
    1. Preparation stage **(P)**, stages before the PSG recording starts
        1. In data_100Hz, the preparation stage **(P)** is labeled as the wake stage **(W)**
    2. Wake Stage **(W)**
    3. Non-rapid eye-movement (NREM) stage 1 **(N1)**
    4. Non-rapid eye-movement (NREM) stage 2 **(N2)**
    5. Non-rapid eye-movement (NREM) stage 3 **(N3)**
    6. Rapid eye movement (REM) **(R)**
    7. No sleep stage labeled on PSG **(Missing)**:
        1. "Missing" label is low. Found only in 2 participants, who had their PSG re-set up during the overnight study, resulting in 15 minutes of consecutive missing labels each. "Missing" label in 4 other participants.

Participant Info:

1. Information Included:
    1. SID (Participant ID), Gender, age, BMI, OAHI, AHI, Arousal Index, Mean_SaSO2, MEDICAL_HISTORY(the previous Medical History), and the Sleep_Disorders (presence or absence of sleep disorders)

Polysomnography (PSG) Signals:

1. **EEG Channels**:
    1. **C4-M1, f4-M1, O2-M1, T3-CZ, CZ-T4**: Channels that recorded brain activity from different locations on the scalp, which provided insights into sleep stages and cortical activity.
2. **EOG Channels**:
    1. **E1, E2**: Channels that monitor eye movements, which are essential for identifying REM sleep.
3. **EMG Channel**:
    1. **CHIN**: Monitors chin muscle activity to detect muscle tone changes during sleep
4. **ECG Channel**:

    1. **ECG**: This records the electrical activity, which is essential for detecting heart rate and rhythm
5. **Respiratory Channels**:
    1. **PTAF**: Airflow monitored by pressure transducer
    2. **FLOW**: Airflow monitored by temperature changes.
    3. **THORAX, ABDOMEN**: Monitors respiratory effort through chest and abdominal movements.
6. **Snoring**:
    1. **SNORE**: monitors snoring sounds.
7. **Leg Movement Channels**:
    1. **LAT, RAT**: Monitors left and right muscle activity for periodic leg movements.
8. **Blood Oxygen Saturation**:
    1. **SAO2**: Monitors the percentage of oxygen in the blood.

B. ***Isaiah Atcherson****:*

Methods **Preprocessing**:

- Removed Unneeded values • **Extracted Features**:
- HR_mean, HR_std, IBI_mean, IBI_std, BVP_mean, BVP_std, Medical_History
- Encoded categorical variables using **BinaryEncoding**:
- Has_Disorder
- Create the Features that go into the dataset:
- Age, BMI, Mean_SaO2, OAHI, AHI, Arousal Index
- Split the dataset into:
- **80% Training**
- **20% Testing**

Modeling Approach

- **Decision Tree Classifier** (scikit-learn)
- Calculated on the physiological metrics taken of all the participants and whether they have a sleep disorder or not
- Trained the model on the training set and made predictions on the test set.

Old Performance Metrics

- Feature Importance
- PCA
- Cross-Validation
- Classification
- Training, Testing Accuracy

New Performance Metrics

- Feature Importance
- Cross-Validation
- Training, Testing Accuracy

Analysis

- Examined model coefficients to understand how strongly each feature influenced the target.
- **Plots/Graphs**
    - o **Decision Trees**
    - o **Heatmap**
    - o **Feature Importance**
    - o **Confusion Matrix**
    - o **Feature Distribution**.

C. ***Bruce Metoyer***

Methods:

Preprocessing:

- Removed missing values
- **Extracted Features**:
    - HR_mean, HR_std, IBI_mean, IBI_std, BVP_mean, BVP_std
- Encoded categorical variables using **Label Encoding**:
    - Sleep_Disorders_Encoded
    - Medical_History_Encoded
- Split the dataset into:
    - **80% Training**
    - **20% Testing**

Modeling Approach

- **Linear Regression** (scikit-learn)
- Calculated mean and standard deviation of the wearable physiological signals during sleep (HR, IBI, BVP).
- Modeled each physiological metric as the target in separate runs.
- Trained the model on the training set and made predictions on the test set.

Performance Metrics

- Mean Squared Error **(MSE)**
- Root Mean Squared Error **(RMSE)**
- Mean Absolute Error **(MAE)**

Analysis

- Examined model coefficients to understand how strongly each feature influenced the target.
- **Plots/Graphs**
    - **Residual plots**: Used to observe the prediction error distribution.

- **Actual vs Predicted plots**: Used to visualize model performance.

D. Muna Chongo

Visualization of Sleep Quality Patterns

To identify patterns associated with low sleep quality, we performed demographic visualizations using the computed Sleep Quality Index (SQI). Participants were grouped by categorical features including gender, obstructive sleep apnea (OSA) status, age, and body mass index (BMI). We employed violin plots to observe the distribution of SQI within each group, and box plots with standard deviation to summarize mean SQI across categories. These visualizations were generated using the seaborn and matplotlib Python libraries. The purpose was to investigate how demographic factors affect sleep quality and to inform the development of more accurate predictive models.

IV. RESULTS

Muna Chongo: Our data was first visualized by merging biometric data from wearable devices with participant data. This required cleaning and aggregating the data, afterwards we computed the Sleep Quality Index (SQI) as the sleep stage count listed in the data(N1, N2, N3, R) divided by the total labeled epochs. Using this derived metric, participants were divided into subgroups by age, BMI, gender, and OSA status. We then employed violin plots, heatmaps, and box plots with standard deviation error bars to visualize both the distribution and central tendency of SQI across these subgroups. These visualizations served to surface potential predictors for modeling and offered early insight into population-level sleep health disparities.

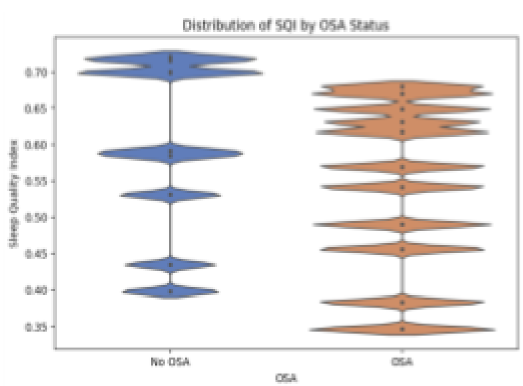


Fig. 1. Violin plot showing SQI distribution by OSA status.

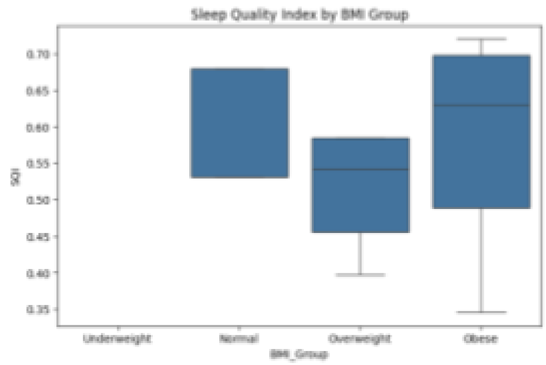


Fig. 2. Box plot of average SQI across BMI categories.

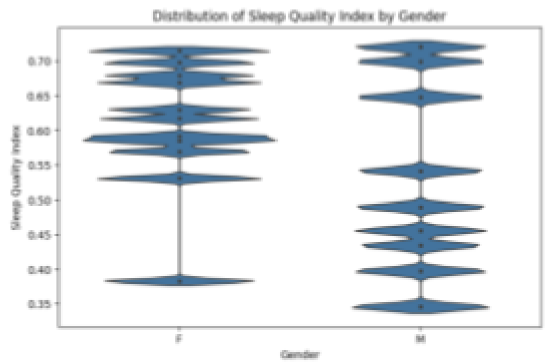


Fig. 3. Violin plot showing SQI distribution by gender.

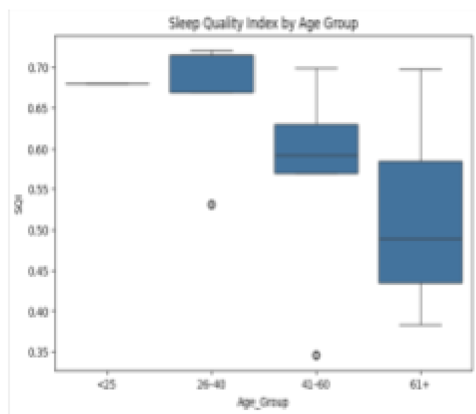


Fig. 4. Box plot showing SQI distribution by Age.

Isaiah Atcherson: The following results will showcase the first code that has been transformed into the second one to fully show the prediction decision tree, and the accuracy of both prediction models. For the first code, it attempted dimension reduction to obtain a better testing accuracy through the linear way of PCA (Principal Components Analysis). For the decision tree, when the leaves are bunched together, it shows that there's too much information in the form of overfitting. There is a heatmap showing how each of the features correlates with each other and with the target. Lastly, there is a distribution chart to show how each feature influences the target and prediction model.

```
Training Accuracy: 1.0
Testing Accuracy: 0.15
Classification Report:
              precision    recall  f1-score   support

           0       0.00      1.00      0.00         0
           1       1.00      1.00      1.00         0
           2       0.50      0.38      0.43         8
           3       0.00      1.00      0.00         0
           4       1.00      1.00      1.00         0
           5       1.00      1.00      1.00         0
           6       0.50      0.33      0.40         9
           7       1.00      1.00      1.00         0
           8       1.00      1.00      1.00         0
           9       0.00      0.00      0.00         5
          10       1.00      1.00      1.00         0
          11       0.00      0.00      0.00         1
          12       0.00      1.00      0.00         0
          13       0.00      0.00      0.00         3
          14       1.00      1.00      1.00         0
          15       1.00      1.00      1.00         0
          16       1.00      1.00      1.00         0
          17       1.00      0.00      0.00         1
          18       0.00      1.00      0.00         0
          19       1.00      1.00      1.00         0
          20       1.00      0.00      0.00         1
          21       0.00      0.00      0.00         3
          22       0.00      1.00      0.00         0
          23       0.00      0.00      0.00         2
          24       1.00      1.00      1.00         0
          25       0.33      0.25      0.29         8
          26       1.00      0.00      0.00         1
          27       0.00      0.00      0.00         1
```

fig 5: Training & Testing Accuracy & Classification Report

The following figure is from the original code, which brings to light the testing and training accuracy, showing that the model wasn't a very good fit, and before the PCA, the testing accuracy was much lower than 15%

```
   micro avg       0.22      0.19      0.20        43
   macro avg       0.55      0.61      0.43        43
weighted avg       0.33      0.19      0.22        43
 samples avg       0.25      0.27      0.24        43

Cross-validation scores: [0. 0. 0. 0. 0.]
Mean cross-validation score: 0.0
Feature Importance:
  feature  importance
1     PC2    0.526598
0     PC1    0.473402
```

fig 6: Feature Importance, Cross-Validation

There was no cross-validation meaning that the model was ill fitted and didn't work fully as intended this would also be shown in Figure 8

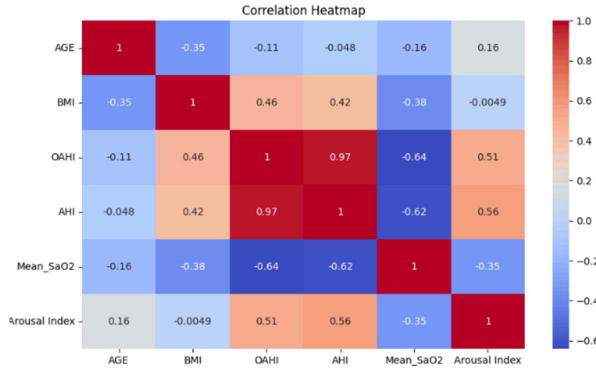


fig 7: Feature Correlation

The features don't correlate well together, other than OAHI and AHI

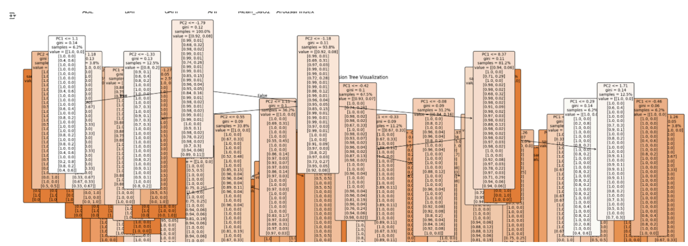


fig 8: Decision tree

The decision tree proves that the model definitely had an overfitting problem, and the values presented were showcasing all the available sleep disorders that could occur in the form of numbers, but that proved to be too much information at once.

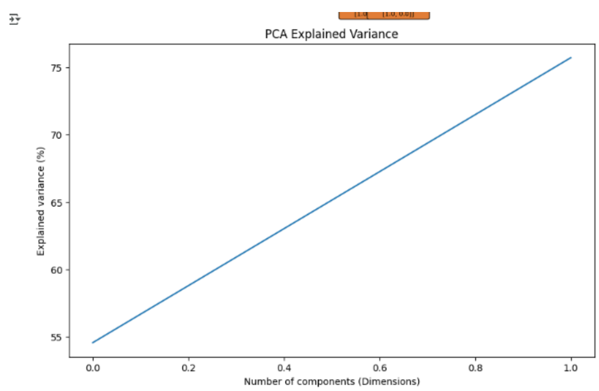


fig 9: PCA Variance

The PCA variance indicates which of the features involved accounts for a proportion of the dataset's variance, and the number of dimensions it passes through

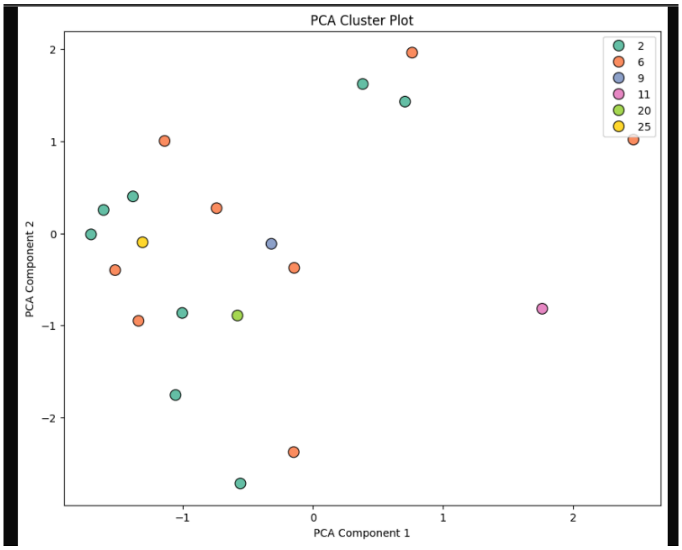


fig 10: PCA Plots

The PCA plots present two differing PCA components that were used to reduce the dimensionality and the correlation between each of the components.

```
Training Accuracy: 0.95
Testing Accuracy: 0.95
Cross-validation scores: [0.9  0.8  0.95 0.9  0.75]
Mean cross-validation score: 0.860

Classification Report:
              precision    recall  f1-score   support

           0       0.00      1.00      0.00         0
           1       1.00      0.95      0.97        20

    accuracy                           0.95        20
   macro avg       0.50      0.97      0.49        20
weighted avg       1.00      0.95      0.97        20


Feature Importances:
        Feature  Importance
1           BMI    0.462428
3          OAHI    0.205740
5 Arousal Index    0.148456
2      Mean_SaO2   0.115607
0           AGE    0.067770
4           AHI    0.000000
```

fig 11: Training & Testing Accuracy + Metrics For Code 2

The training and testing accuracy are closely related in this upgraded code, and the features are shown in terms of importance.

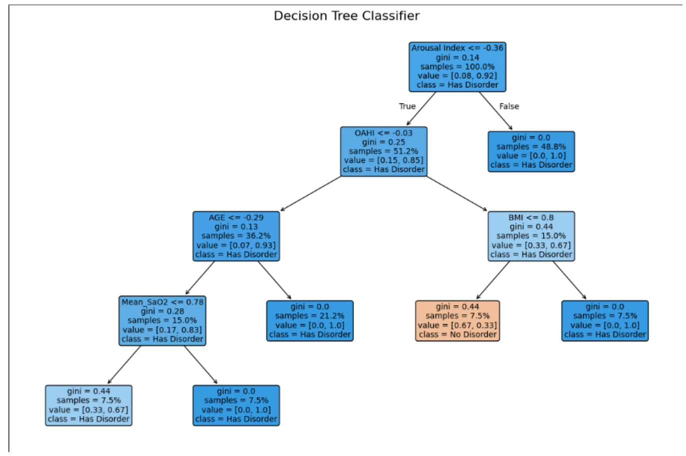


fig 12: New Decision Tree

A smaller Decision Tree is shown to help with the model visualization and the model predictions

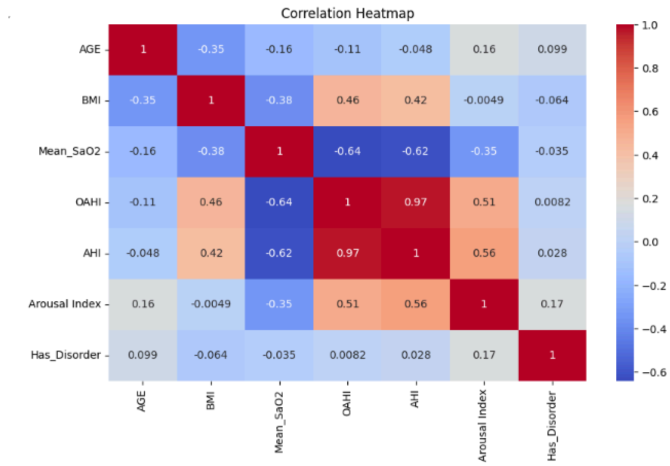


fig 13: Correlation Plot

New correlation plot that also features the correlation between the features and if the user will have a disorder or not

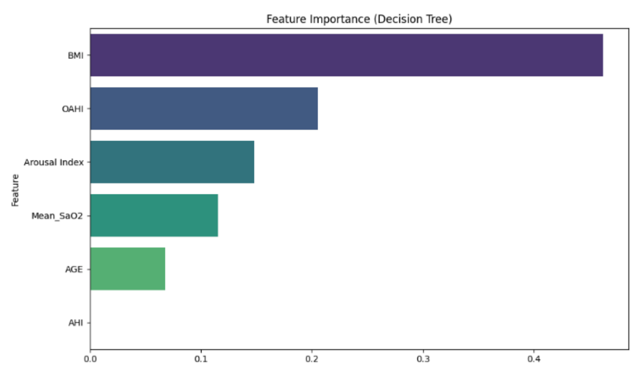


fig 14: Feature Importance

The barplot signifies which of the features was the best associated fit for the decision tree and the prediction plot

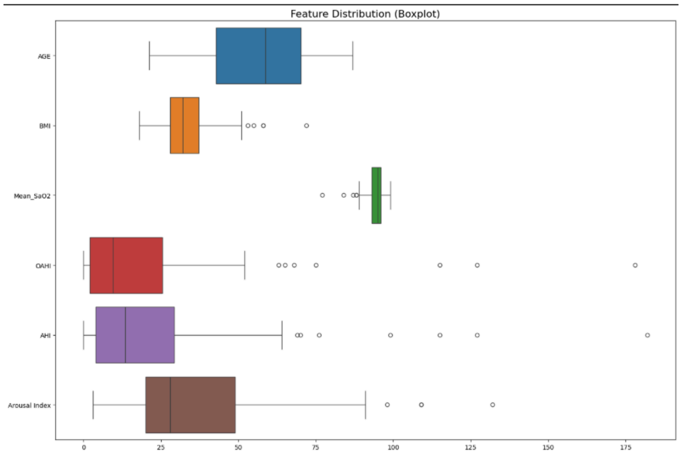
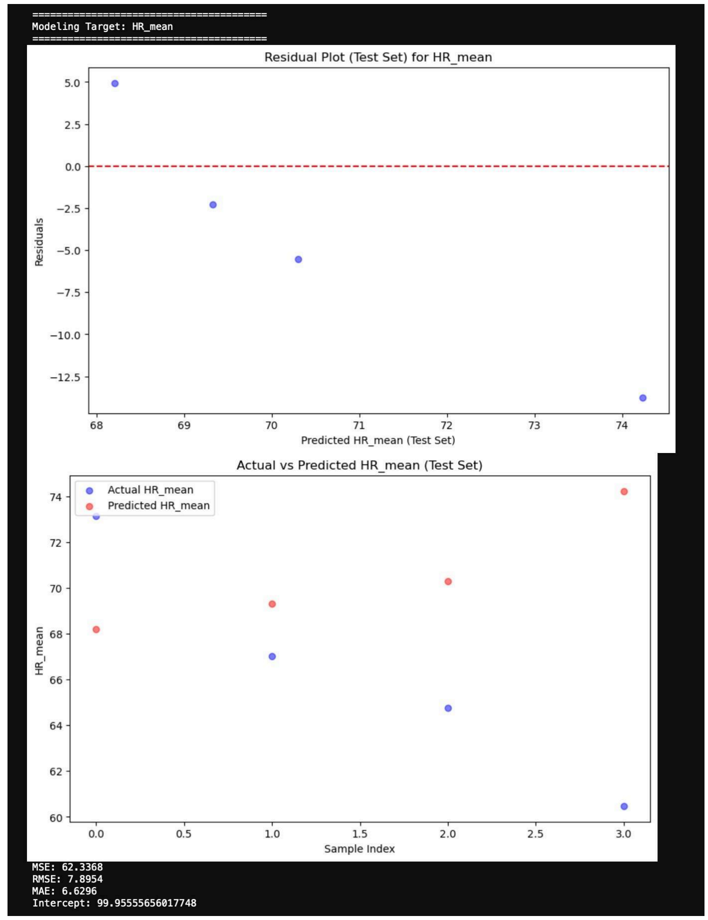


fig 15: Feature Distribution

Lastly the distribution of the features shows which fit into the model. The outliers signify the features such as BMI who is underweight or overweight, or as in the case of age which is too old or too young for the prediction model, though it was still used in the model regardless.

Bruce Metoyer: The results will include residual plots, as well as actual versus predicted plots. The residual plots visualize the accuracy of the predictions about the target variable. The residuals (plot points) are randomly scattered around zero, so the model is performing well. On the other hand, if the plot points are scattered, then the model isn't performing well. When examining the actual vs. predicted plot, we are evaluating how closely the model's predictions align with the actual observations. The closer the red and blue points align, the better the model's performance. Lastly, the Coefficient chart shows the influence of each feature. Positive correlations indicate a direct relationship, while negative ones suggest an inverse correlation.



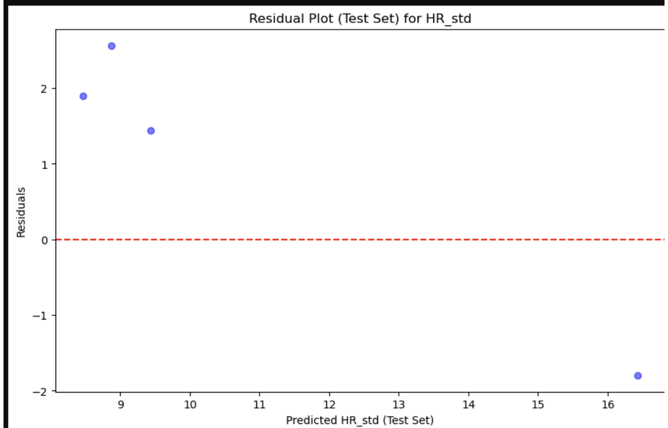
The HR_mean of the residual plot is linearly related to physiological variables, such as sleep disorder status, while the actual vs. predicted HR_mean has decent performance.

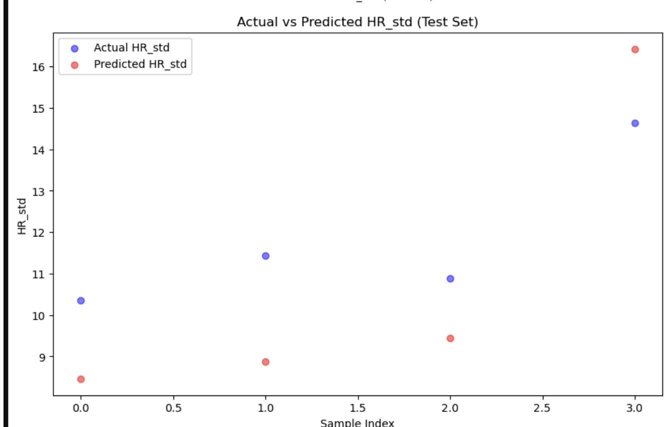
Coefficients DataFrame:

	Estimated Coefficients
HR_std	0.022715
IBI_mean	-52.474460
IBI_std	68.836368
BVP_mean	-870.917259
BVP_std	0.067660
Sleep_Disorders_Encoded	1.863061
Medical_History_Encoded	-0.053393

The Coefficients of HR_Mean show BVP_mean, IBI_mean, and the medical histories to have the worst relationship regarding HR_Mean, while IBI, BVP_std, and sleep disorders have a positive relationship.

```
Modeling Target: HR_std
========================================
```

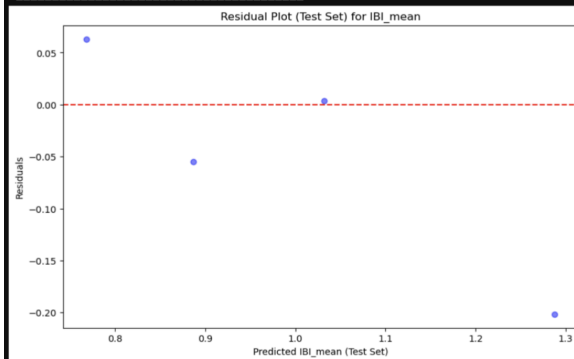


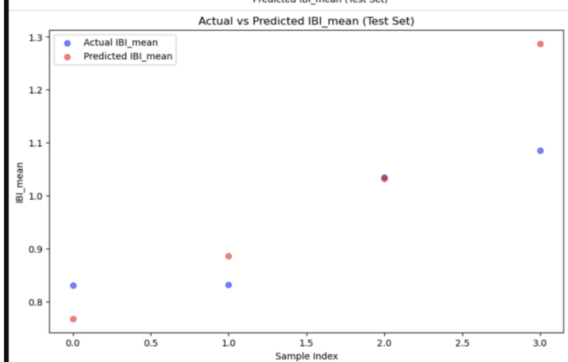



```
MSE: 3.8592
RMSE: 1.9645
MAE: 1.9219
Intercept: 0.6695467759361531

Coefficients DataFrame:
                          Estimated Coefficients
HR_mean                                 0.008870
IBI_mean                               -0.215422
IBI_std                                53.072645
BVP_mean                             -130.924083
BVP_std                                 0.042701
Sleep_Disorders_Encoded                 0.366852
Medical_History_Encoded                -0.103317
```

The standard deviation of HR shows similar results to its mean regarding the residual plot; however, the actual vs. predicted plot exhibits a more linear relationship. The coefficients reveal that IBI standard deviation has the strongest positive relationship, while the mean of BVP has the weakest relationship.

```
Modeling Target: IBI_mean
========================================
```

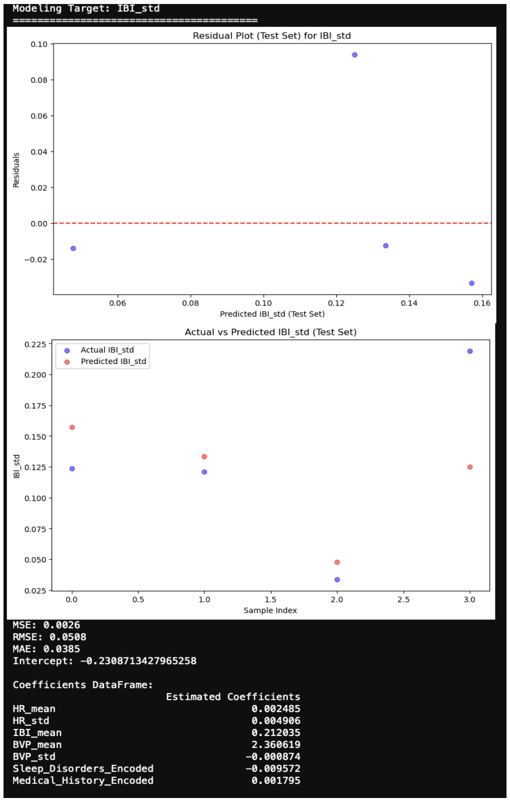


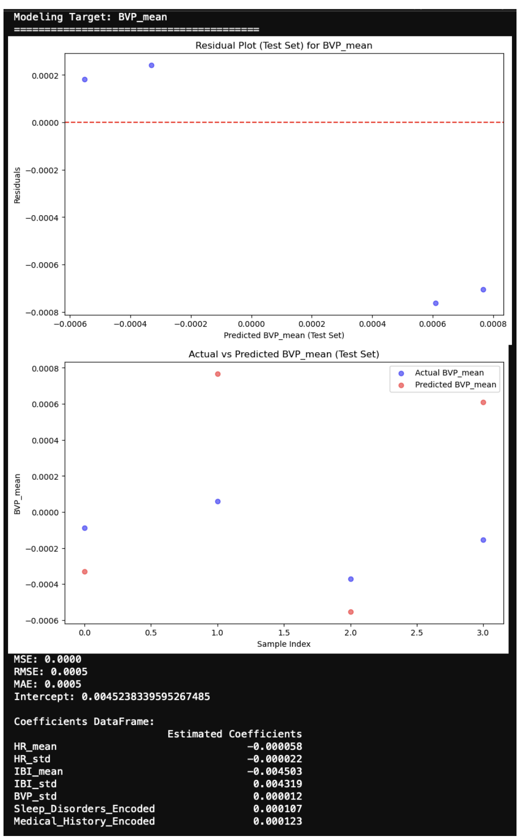


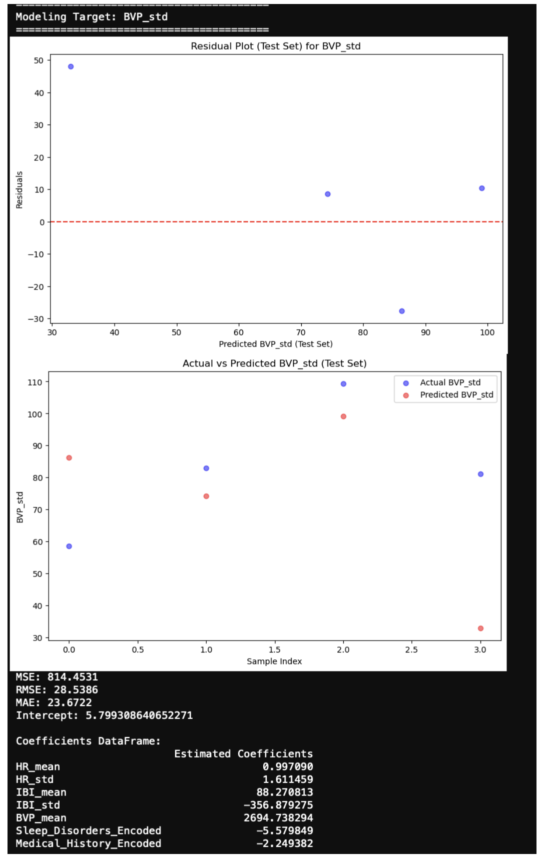

```
MSE: 0.0119
RMSE: 0.1091
MAE: 0.0807
Intercept: 1.5353778746918254

Coefficients DataFrame:
                          Estimated Coefficients
HR_mean                                -0.014385
HR_std                                 -0.000151
IBI_std                                 1.610333
BVP_mean                              -18.688504
BVP_std                                 0.001642
Sleep_Disorders_Encoded                 0.024355
Medical_History_Encoded                -0.000604
```

The IBI_mean and IBI_std for both the residual plot and actual vs predicted values showed the strongest correlation in both plots. For IBI_mean, there was a strong negative coefficient relationship with BVP_mean, with a moderate positive coefficient relationship with IBI_std. When examining IBI_std coefficients, we observe multiple moderate positive relationships, with BVP_mean being the strongest, while little to no negative coefficients were found.

Lastly, BVP_mean and BVP_std yielded nearly identical results to those of HR; however, we see a stronger actual vs predicted relationship between BVP_std. The coefficients for BVP_mean had little to no influence on the results. When looking at BVP_std, there is a strong BVP_mean with 2694 and 88 for IBI_mean, with a strong negative relationship in IBI_std.







V. DISCUSSION

Bruce Metoyer:

Residual Plot and Actual vs Predicted Plots:

- HR_Mean
  - The residuals showed a somewhat weak pattern, indicating a relatively unbiased model. This reveals that HR_mean is somewhat linearly related to physiological variables, such as sleep disorder status, supporting the notion that heart rate has an impact on sleep quality.
- HR_std
  - HR_std is greater than the variability and affects the heart rate. HR_std showed similar results to HR_mean, indicating that fluctuations in heart rate reflect a stronger relationship with sleep quality, especially when analyzing the actual vs. predicted plot points.
- IBI_Mean & IBI_std
  - Both the IBI mean and standard deviations showed the strongest relationship among all of the metrics. IBI reflects heart rhythm

- variability and average heart rate, indicating that strong relationships have a greater impact on overall sleep quality.
- BVP_Mean & BVP_std
  - Both standard deviation and mean revealed stronger results when compared to HR, but not as strong as IBI. The variation and variability in blood volume reflect the relationship with sleep quality. Blood volume is connected to not only sleep quality but also possibly blood pressure and the effects that different foods can have on your health.

Coefficient Chart:

- Out of all of the Coefficients, BVP_mean revealed the strongest relationship with a positive value of 2694.738 on BVP_std. This makes the most sense, as both the BVP mean and the standard deviation of blood volume heavily influence each other. Sticking with BVP_std, it also produced the strongest negative relationship with IBI_std. This relationship suggests that variability in heart rhythm is more closely linked to blood volume. This reveals that blood flow is reduced at an increased heart rate and vice versa.

Isaiah Atcherson: The results showed that there are correlations between age, BMI, arousal index, and Mean_SaO2, which affect a person's ability to develop sleep disorders. As seen in Figure 13, there is some slight correlation. Both of the Decision trees are used to predict if a person can develop sleep disorders, however, the first one showed too much information, while the second one was just right in terms of leaves and nodes shown. Figure 11 and Figure 5 are images of the testing and training accuracy of the model, and while the training accuracy is similar, the testing is not, even though PCA was used on Figure 5 to help improve it, there is still a significant difference. Though the testing accuracy shown in Figure 11 is much better and is exactly the same as the training, meaning that the model is performing quite well.

Muna Chongo: The visualizations revealed clear patterns that contribute to our understanding of sleep quality differences among patient subgroups. In Fig. 1, participants with OSA exhibited both lower mean SQI and higher variability compared to non-OSA individuals. Fig. 2 shows that BMI was inversely related to sleep quality, particularly in the obese category. Interestingly, the violin plots by gender (Fig. 3) indicated relatively uniform distributions, suggesting that gender may not be a significant standalone predictor of sleep quality in this dataset. Lastly, Fig. 4 demonstrates a decline in average SQI in older age groups, particularly among those aged 60 and above. These findings suggest that demographic and biometric features may enhance the performance of machine learning models by providing contextual sleep-related risk indicators.

VI. CONCLUSION AND FUTURE WORKS

Our visual analysis provided critical insight into how SQI differs across patient subgroups, supporting our goal of developing predictive models that consider individual variability. These insights form a foundation for future models that could offer patient-specific sleep interventions based on physiological and demographic profiles. Through a combination of ML models and demographic analysis, we found that IBI and BVP variability were the strongest indicators of sleep quality, with heart rate variability also showing meaningful correlations. Visualizations and regression analyses also revealed that participants with sleep disorders, a higher BMI, or an older age showed to have poor sleep quality. Our results provide a strong foundation for developing personalized sleep-monitoring tools. Future work should include using more patients in the dataset for more accurate models. Other considerations could include using other advanced machine learning models and analyzing how additional factors can impact an individual's sleep, such as occupation and eating habits.

ACKNOWLEDGMENT

We want to thank the Department of Computer Science at Bowie State University for facilitating this course. This course allowed us to learn about Data Science and ML techniques vital for understanding what it takes to be a data scientist.

REFERENCES

1. [1] K. Wang, J. Yang, A. Shetty, and J. Dunn, “DREAMT: Dataset for Real-time sleep stage EstimAtion using Multisensor wearable Technology,” *Physionet.org*, Feb. 05, 2025. https://physionet.org/content/dreamt/2.0.0/#files-panel (accessed March 20, 2025).