Integrated Health Monitoring Using Wearable Devices

Students: Maria Sanchez and Rafael Linarez
Mentor: Dr. Agoritsa Polyzou

Github: https://github.com/rafaelinarez/Capstone-Wearables-Health-Monitoring

Summary

Wearable gadgets produce great amounts of data related to physical activity, heart rate, METs, and sleep patterns. However, converting this data into practical insights is still a significant challenge. This final project aims to address this issue by creating machine learning models for predicting sleep results and using clustering methods to reveal behavioral trends. Through the application of sophisticated statistical and machine learning techniques, this initiative will investigate how activity, heart rate, and sleep metrics are interconnected to offer a thorough comprehension of user behaviors and potential health advice.

The project will make use of an extensive dataset from wearable technology, comprising minute-by-minute activity logs, heart rate information, and sleep diaries, to build predictive models and clustering algorithms. The outcomes will not only assess the precision of different machine learning approaches but also evaluate the potential of clustering-driven insights in improving customization within wearable devices and health-related software.

1. Introduction

Wearable devices have gained widespread popularity for tracking physical activity, sleep patterns, and heart rate, providing individuals with insights into their health. However, the vast amount of data generated by these devices remains underutilized, especially in delivering actionable insights and personalized recommendations. This gap underscores the need for integrating wearable data with advanced machine learning and statistical techniques to uncover meaningful relationships between activity, sleep, and physiological metrics.

1.1 Hypothesis

This project is guided by the hypothesis that daily physical activity significantly impacts sleep quality and quantity. Specifically:

- **Increased activity** is expected to lead to better sleep outcomes, including falling asleep faster and achieving more restful sleep.
- **Sedentary behavior** is expected to result in poorer sleep quality, characterized by longer durations and disrupted patterns.

These assumptions drive the project's focus on clustering and predictive modeling to validate and explore these interrelationships.

2. Prior Art & Challenges

Research into the interplay between physical activity, sleep, and heart rate (HR) has established foundational insights into health monitoring and predictive modeling. Key studies have demonstrated the potential of wearable device data for estimating energy expenditure, understanding sleep behaviors, and analyzing activity patterns:

1. **Heart Rate as a Predictor of Energy Expenditure:** Hiilloskorpi et al. (Year) highlighted HR's reliability as a predictor of energy expenditure (EE) across various activity intensities using regression models. This work laid the groundwork for using HR as a key variable in energy monitoring systems and supports the inclusion of HR metrics in this project's predictive models [1].
2. **Physical Activity and Sleep Patterns:** McClain et al. (2014) demonstrated positive correlations between moderate-to-vigorous activity and longer sleep durations. These findings affirm the value of integrating activity metrics into sleep outcome models, guiding this project's focus on predicting sleep patterns [2].
3. **Bidirectional Effects Between Activity and Sleep:** Liao et al. (2020) explored the dynamic relationships between daily activity and sleep using wearable data, revealing the influence of sedentary behavior on sleep duration. This study informs this project's temporal analysis approach for modeling interdependencies between these variables [3].
4. **Integrating Heart Rate and Acceleration:** Nakanishi et al. (2018) showcased the benefits of combining HR and acceleration data to improve MET estimation, emphasizing multi-modal approaches for predictive accuracy. This integration aligns with the project's objective to combine activity, HR, and sleep metrics for comprehensive modeling [4].
5. **Predicting Sleep Using Wearables and Machine Learning:** Park et al. (2024) proposed a model using wearable devices to predict sleep outcomes by combining physical activity, light exposure, and heart rate variability (HRV). Their study demonstrated the utility of HRV as a psychological stress indicator, significantly enhancing predictive accuracy (XGBoost achieved 85% accuracy). The work underscores the potential of multi-modal data and machine learning for advancing sleep prediction, aligning closely with this project's objectives [5].

Challenges

While prior studies have laid the groundwork for activity, HR, and sleep analysis, significant challenges remain:

- **Missing Data:**

    - Sleep and heart rate records often have high levels of missingness (e.g., >50% in some metrics in this dataset). Ensuring reliable predictions and imputation of missing values remains a critical challenge.
    - Handling user-level variability in missing data across metrics (e.g., some users lack sleep data entirely) complicates modeling and can lead to biased insights.

- **Data Integration:**
    - Combining metrics from diverse categories (e.g., HR, activity, and sleep) introduces complexity in feature engineering and necessitates robust techniques to capture meaningful relationships across variables.
    - Temporal alignment of features across minute-level and daily-level data requires careful preprocessing to ensure consistency.

- **Model Selection and Generalization:**
    - Evaluating a wide range of machine learning models (regression, SVR, KNN, XGBoost) for predictive tasks demands balancing accuracy and computational efficiency.
    - Generalizing insights from a relatively small dataset (33 participants) poses a challenge for ensuring broader applicability.

- **Clustering and Classification Challenges:**
    - Using clustering techniques (e.g., k-means) to classify users as “Active” or “Sedentary” can suffer from overlapping clusters and low inter-cluster variability.
    - Predicting user clusters using sleep data alone may result in moderate accuracy, requiring exploration of additional features or advanced clustering algorithms.

- **Interpretability vs. Complexity:**
    - Balancing interpretability (e.g., linear regression) and predictive power (e.g., gradient boosting) is essential to generate actionable insights for end users of wearable technology.

How the Project Addresses These Challenges

- Advanced Data Imputation: Leveraging tailored techniques (e.g., gradient boosting for missing HR data, regression models for sleep metrics) ensures data completeness and reduces bias.
- Robust Modeling: Comparing linear and nonlinear models helps identify trade-offs between simplicity and accuracy, while clustering enhances understanding of user behaviors.
- Cross-Validation and Evaluation: Employing rigorous cross-validation and metrics (e.g., RMSE, Silhouette Score) ensures model reliability and generalizability.

- Insights Generation: Statistical tests (e.g., ANOVA) and clustering analyses provide actionable insights into the relationships between activity, HR, and sleep metrics.

3. Objectives

The primary objective of this project is to analyze wearable data to uncover actionable health insights and provide personalized recommendations. To achieve this, the project focuses on the following key goals:

- **Develop Predictive Models:**
    - Build and fine-tune machine learning models to forecast sleep outcomes (Total Sleep Records, Total Minutes Asleep, and Total Time in Bed) using activity levels, METs (metabolic equivalent of task), and heart rate data.
    - Evaluate the performance of these models using metrics such as RMSE, precision, and recall.

- **Use Clustering to Classify Users:**
    - Apply K-Means clustering to group users based on activity and physiological metrics.
    - Analyze clusters to identify common patterns in behavior and sleep outcomes.
    - Validate clustering quality using metrics such as the Silhouette Score and PCA visualizations.

- **Evaluate Machine Learning Techniques:**
    - Compare the effectiveness of various machine learning models, including linear and non-linear approaches, to identify the most suitable model for sleep outcome prediction.

- **Provide Personalized Health Recommendations:**
    - Use insights from predictive modeling and clustering analysis to offer actionable recommendations that optimize user activity and sleep patterns.

4. Motivation

Sleep is a critical determinant of overall health and well-being, yet many wearable devices lack the ability to translate data into meaningful insights for improving sleep patterns. Understanding the interplay between activity, heart rate, and sleep behaviors offers significant potential for enhancing personalized health recommendations.
This project is motivated by:

- **Health Benefits:** Providing users with actionable insights to optimize their sleep and activity patterns, contributing to improved health outcomes.
- **Advancing Wearable Technology:** Enhancing the utility of wearable devices by moving beyond simple data tracking to generating predictive and prescriptive insights.
- **Addressing Data Challenges:** Tackling issues such as missing data and variability in wearable datasets to improve analytical reliability.
- **Real-World Impact:** Informing the design of smarter wearable algorithms and user interfaces that can dynamically adapt to individual behaviors and needs.

This capstone project serves as a proof-of-concept for leveraging wearable data to uncover valuable insights and build models that promote better health through data-driven personalization.

5. Data Sources and Description

The dataset utilized in this project was sourced from Kaggle, containing minute-level data for 33 Fitbit users over a span of two months (March 12, 2016, to May 12, 2016). Each record represents a day of activity for an individual, capturing a comprehensive range of metrics such as physical activity, sleep patterns, heart rate, and energy expenditure (measured via METs).

Key Characteristics:

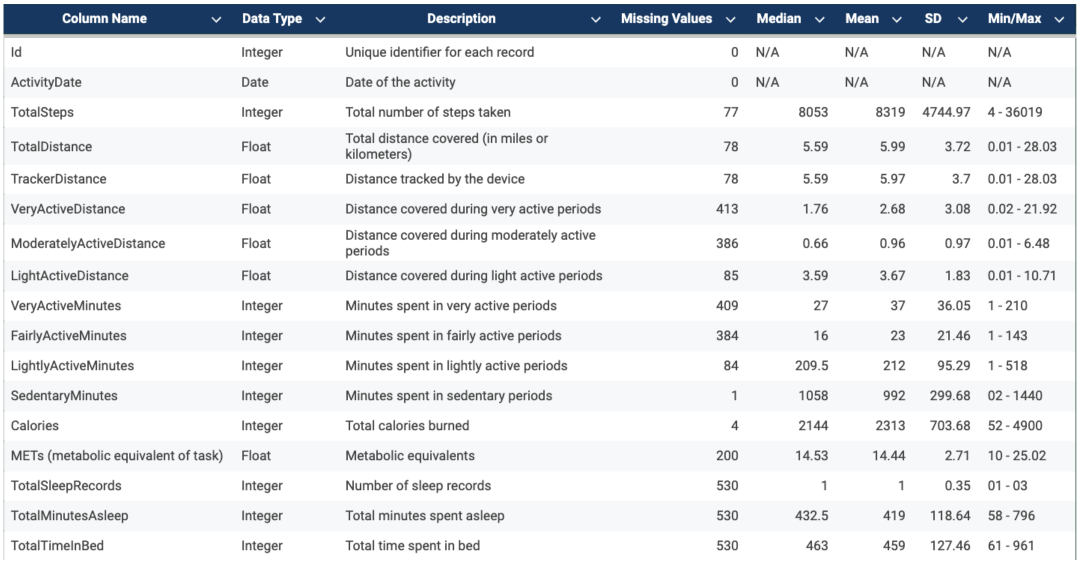
1. **Users and Records:**
    - Total Users: 33 unique individuals identified by a unique Id.
    - Average Records per User: Approximately 28.49 entries per user.
    - Data coverage varies across categories, with some users having complete records while others display significant gaps.

2. **Key Data Columns:**
    - Activity Metrics:
        - Total Steps: Represents daily step count, with values ranging from 4 to 36,019 steps. There are 77 missing entries across users.
    - Sleep Metrics:
        - Total Minutes Asleep: Captures sleep duration, with a mean of 419 minutes and significant missing data (56.3%).
        - Total Time in Bed: Tracks time spent in bed, including sleep and non-sleep activities. This metric also has 56.3% missing entries.
    - Heart Rate Metrics:
        - Heart Rate (Max, Min, Avg): These metrics often have high missing rates, varying significantly across users.
    - Energy Expenditure Metrics:
        - METs (Metabolic Equivalents): Reflects energy spent on physical activities. Approximately 200 entries are missing.

3. **Missing Data Analysis:**

- Some users have complete records across all categories, while others show high proportions of missing data, particularly for sleep and heart rate:
    - 7 users: Have complete records across all categories.
    - 14 users: Are missing data in one category.
    - 12 users: Have missing data in two or more categories.
- Specific examples of users with significant missing data:
    - User 1624580081: Missing all sleep and heart rate data (31 records each).
    - User 4020332650: Missing 14 activity records, 23 sleep records, 15 heart rate records, and 1 MET record.

4. **Data Quality Improvements:**
    - To enhance the quality and completeness of the dataset:
        - Records with full-day sedentary activity (e.g., SedentaryMinutes = 1440) were removed.
        - Users with fewer than 15 days of activity data were excluded.
        - Data across categories were merged using unique identifiers (Id and ActivityDate) for unified analysis.



Column Name	Data Type	Description	Missing Values	Median	Mean	SD	Min/Max
Id	Integer	Unique identifier for each record	0	N/A	N/A	N/A	N/A
ActivityDate	Date	Date of the activity	0	N/A	N/A	N/A	N/A
TotalSteps	Integer	Total number of steps taken	77	8053	8319	4744.97	4 - 36019
TotalDistance	Float	Total distance covered (in miles or kilometers)	78	5.59	5.99	3.72	0.01 - 28.03
TrackerDistance	Float	Distance tracked by the device	78	5.59	5.97	3.7	0.01 - 28.03
VeryActiveDistance	Float	Distance covered during very active periods	413	1.76	2.68	3.08	0.02 - 21.92
ModeratelyActiveDistance	Float	Distance covered during moderately active periods	386	0.66	0.96	0.97	0.01 - 6.48
LightActiveDistance	Float	Distance covered during light active periods	85	3.59	3.67	1.83	0.01 - 10.71
VeryActiveMinutes	Integer	Minutes spent in very active periods	409	27	37	36.05	1 - 210
FairlyActiveMinutes	Integer	Minutes spent in fairly active periods	384	16	23	21.46	1 - 143
LightlyActiveMinutes	Integer	Minutes spent in lightly active periods	84	209.5	212	95.29	1 - 518
SedentaryMinutes	Integer	Minutes spent in sedentary periods	1	1058	992	299.68	02 - 1440
Calories	Integer	Total calories burned	4	2144	2313	703.68	52 - 4900
METs (metabolic equivalent of task)	Float	Metabolic equivalents	200	14.53	14.44	2.71	10 - 25.02
TotalSleepRecords	Integer	Number of sleep records	530	1	1	0.35	01 - 03
TotalMinutesAsleep	Integer	Total minutes spent asleep	530	432.5	419	118.64	58 - 796
TotalTimeInBed	Integer	Total time spent in bed	530	463	459	127.46	61 - 961

Relevance to the Project

This dataset provides a rich foundation for understanding relationships between physical activity, sleep behavior, and health metrics. The extensive variability and missing data necessitated careful preprocessing, such as data imputation and cleaning, to ensure the reliability of clustering and predictive modeling outcomes.

6. Methods and Tools

Overview

This project was structured into three main phases to analyze data from wearable devices and derive actionable health insights:

1. **Phase 1: Data Loading and Cleaning**
   - This phase focused on preparing the dataset for analysis. Data from various sources was merged, missing values were handled using robust imputation techniques, and exploratory data analysis (EDA) was conducted to ensure readiness for subsequent modeling.

2. **Phase 2: Clustering Analysis**
   - To answer the question, *"How can clustering based on sleep patterns and activity levels identify distinct groups of users, and what insights can these clusters provide about variations in sleep behavior and activity interactions?"*, K-Means clustering was applied. The analysis helped identify distinct behavioral patterns and user groupings.

3. **Phase 3: Predictive Modeling**
   - To address the question, *"How accurately can sleep outcomes (TotalSleepRecords, TotalMinutesAsleep, TotalTimeInBed) be predicted using activity levels, METs, and heart rate data?"*, predictive models were developed using various machine learning algorithms. This step aimed to provide personalized sleep recommendations and evaluate the effectiveness of predictive techniques.

The following sections describe the tools and techniques applied across these phases.

6.1 Tools and Libraries

The project leveraged a variety of Python libraries and tools to process and analyze the data effectively. Below are the key libraries and their applications:

- **Pandas**: For data manipulation, cleaning, and integration, particularly for handling missing values and merging datasets from multiple sources.
- **NumPy**: Used for numerical computations and feature engineering, such as aggregation and handling missing data.
- **Matplotlib and Seaborn**: To create visualizations for exploring data distributions, identifying patterns, and interpreting results from clustering and predictive modeling.
- **Scikit-learn**: A versatile library used for:
  - Data imputation techniques like K-Nearest Neighbors (KNN).
  - Clustering analysis (e.g., K-Means) to identify user groups.
  - Predictive modeling using regression algorithms like Ridge, Lasso, and Elastic Net.

    - Hyperparameter tuning and cross-validation for optimizing model performance.
- **Statsmodels and Scipy**: Employed for statistical tests, including Shapiro-Wilk tests for normality, and validating model assumptions.
- **XGBoost**: For high-performance gradient boosting in predictive modeling, particularly in forecasting sleep metrics.

These tools were selected to address the challenges of cleaning, integrating, and analyzing a dataset with missing data and multiple metrics. They were applied strategically to ensure robust preprocessing, insightful clustering, and accurate predictive modeling. Each phase demonstrates the practical application of these tools in detail.

6.2 Phase 1: Clustering

The primary goal of this phase was to address the question: *"How can clustering based on sleep patterns and activity levels identify distinct groups of users, and what insights can these clusters provide about variations in sleep behavior and activity interactions?"* The purpose of clustering was to identify user behavior patterns and group users based on activity and sleep metrics to uncover actionable insights.

6.2.1 Methodology

- **Clustering Approach**
    - **K-Means Clustering** was selected due to its simplicity and effectiveness for behavioral segmentation.
    - Two approaches were implemented:
        1. **Clustering on All Features**: Applied K-Means using the full feature set, excluding sleep data.
        2. **Clustering on Selected Features**: Features with significant inter-cluster differences (e.g., total steps, sedentary minutes) were selected to refine the clustering process.

- **Parameters and Evaluation Metrics**
    - **K-Means Settings**:
        1. Initialization: 'k-means++' to improve convergence.
        2. Number of Clusters (k): 2 (based on preliminary analysis).
        3. n_init: 10 (number of runs for stability).
        4. Random State: 100 (for reproducibility).
    - **Evaluation Metrics**:
        1. **Silhouette Score**: Measures cluster cohesion and separation.
        2. **Inertia**: Quantifies the sum of squared distances between data points and their assigned cluster centers.

- **Feature Selection and Refinement**

- Features with minimal differences between clusters were excluded (e.g., light active minutes).
- Significant features (e.g., total steps, sedentary minutes) were retained to improve cluster definition.

6.2.2 Results

1. **Silhouette Score Improvement**:
    - Clustering on all features achieved a **Silhouette Score of 0.2464**.
    - Clustering on selected significant features improved the score to **0.33**, indicating better-defined clusters.
    -

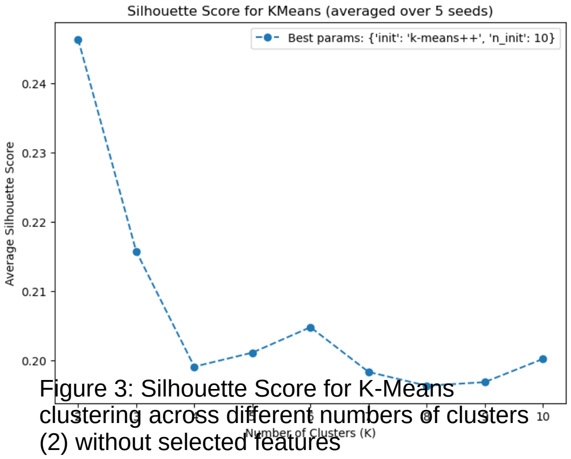
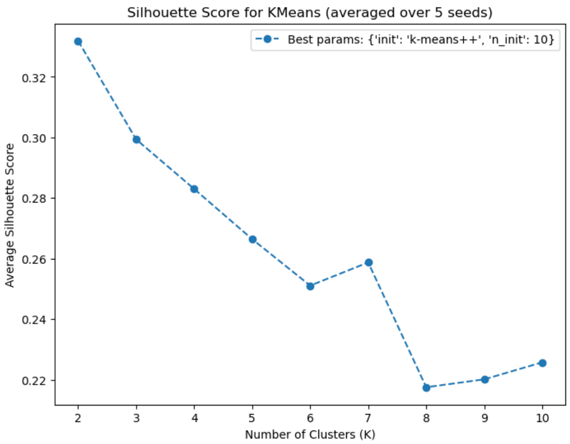


Figure 3: Silhouette Score for K-Means clustering across different numbers of clusters (2) without selected features



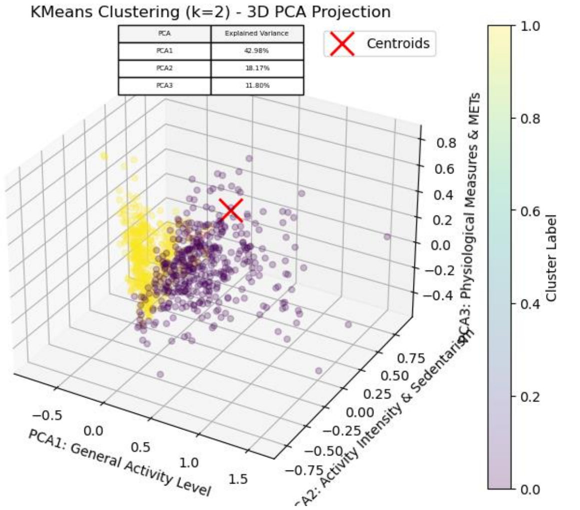
selected features

The Silhouette Score graphs above illustrate the clustering performance across different numbers of clusters. With feature selection, the score improved consistently, reaching a peak at k=2 (0.33), reflecting better-defined clusters compared to using all features.

2. **Cluster Insights**:
    - **Cluster 0**: Represented users with consistent, moderate activity patterns and higher total steps.
    - **Cluster 1**: Represented users with patterns dominated by sedentary minutes and lower overall activity.

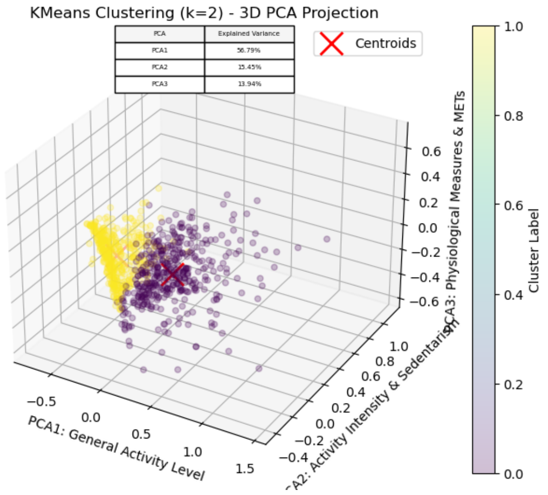
3. **Behavioral Patterns**:
    - Users in Cluster 0 exhibited more active lifestyles, associated with potentially better sleep outcomes.
    - Cluster 1 users displayed higher sedentary behavior, which aligns with poorer sleep quality, consistent with project assumptions.

The PCA visualizations below provide a comparison of clustering results before and after feature selection. The refined approach demonstrates improved cluster separation, reducing overlap and enhancing interpretability.



PCA	Explained Variance
PCA1	42.98%
PCA2	18.17%
PCA3	11.80%

Figure 5: 3D PCA visualization of K-Means clustering using all features, showing moderate cluster separation (Silhouette Score = 0.2464)



PCA	Explained Variance
PCA1	56.79%
PCA2	15.45%
PCA3	13.94%

Figure 6: 3D PCA visualization of K-Means clustering with selected features, demonstrating improvement cluster separation (Silhouette Score = 0.33)

4. **Validation**:
   - Statistical validation was performed using ANOVA to confirm significant differences in key metrics (e.g., total minutes asleep, time in bed) between clusters.
   - Results demonstrated low p-values (<0.05), affirming the effectiveness of clustering in segmenting users based on behavioral patterns.

5. **KNN Model Performance:**

   Model Setup:

- The KNN classifier was trained to predict cluster assignments using sleep records (Total Sleep Records, Minutes Asleep, and Time in Bed) as features.
- Target variables were defined as:
  - Class 0: Active_Users (higher activity).
  - Class 1: Sedentary_Users (lower activity).
- Best hyperparameters were determined based on validation error:
  - metric='manhattan', n_neighbors=5, weights='uniform'.

- **Performance Metrics:**
  - **Class Distribution in Test Set:**
    - Class 0 (Active_Users): 55 samples
    - Class 1 (Sedentary_Users): 52 samples
  - **Confusion Matrix Analysis:**

- **Class 0 (Active Users):**
    - Precision: 66%
    - Recall: 69%
    - F1-Score: 67%
- **Class 1 (Sedentary Users):**
    - Precision: 65%
    - Recall: 62%
    - F1-Score: 63%
- **Overall Accuracy:** 65%

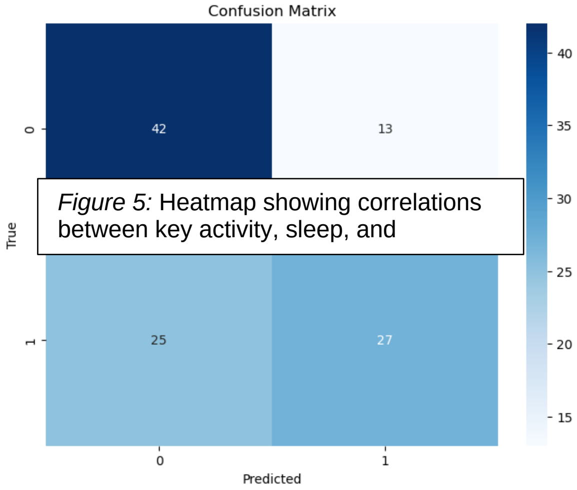


Figure 5: Heatmap showing correlations between key activity, sleep, and

Insights:

- Accuracy of 65% indicates reasonable performance in predicting clusters based on sleep data.
- Cluster-specific performance metrics reveal better prediction accuracy for Class 0, likely due to less overlap in features compared to Class 1.
- Practical applications include the ability to classify new users and provide tailored activity or sleep recommendations.

Heatmap Analysis: The heatmap illustrates the relationships between activity, sleep, and heart rate metrics across the dataset. Key insights include:

- Strong positive correlations between Total Steps and Total Distance.
- Weak or negative correlations between sedentary behavior metrics and sleep metrics, supporting the hypothesis that higher activity levels are associated with better sleep outcomes.

6.3 Phase 2: Data Cleaning and Preprocessing

The primary goal of this phase was to prepare the dataset for analysis by:

- Integrating data from multiple sources into a unified dataset.
- Addressing missing data using robust imputation techniques.
- Conducting exploratory data analysis (EDA) to understand data distributions, identify outliers, and ensure readiness for subsequent modeling.

6.3.1 Steps:

1. **Data Loading and Integration**

    - **Data Sources:** The data consisted of daily activity, heart rate, metabolic equivalents (METs), and sleep metrics sourced from multiple CSV files.
    - **Standardization:**
        - Dates were standardized across all datasets to ensure consistent temporal alignment.
        - Columns were renamed and reformatted as needed for uniformity.

- **Aggregations:**
    - Daily summaries of heart rate (max, min, and average) and METs (max, min, and average) were computed.
- **Merging:**
    - The datasets were merged using unique identifiers (Id) and daily activity dates (ActivityDate), creating a unified dataset with all relevant metrics.

2. **Handling Missing Data**

    - **Missingness Identification:**
        - Significant missing values were observed in sleep and heart rate metrics, requiring careful imputation.
    - **Imputation Techniques:**
        - Statistical imputation (mean/median) was applied based on feature distributions (verified using the Shapiro-Wilk test).
        - K-Nearest Neighbors (KNN) imputation was used for features with complex dependencies, ensuring accurate reconstruction of missing values.

3. **Data Cleaning**

    - Records with fully sedentary days (SedentaryMinutes = 1440) were removed to exclude non-representative data points.
    - Users with fewer than 15 days of activity data (less than 50% completeness) were excluded.
    - Duplicate records were checked and removed to ensure data integrity.

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

    - **Descriptive Statistics:**
        - Summary statistics were computed for key metrics such as TotalMinutesAsleep, TotalSteps, and heart rate measures.

The heatmap below illustrates the correlations between key metrics in the dataset, such as activity levels, heart rate values, and sedentary minutes. Strong correlations, such as those between TotalSteps and TotalDistance, highlight expected relationships, while weaker correlations, such as those involving SedentaryMinutes, suggest independent behavioral patterns.

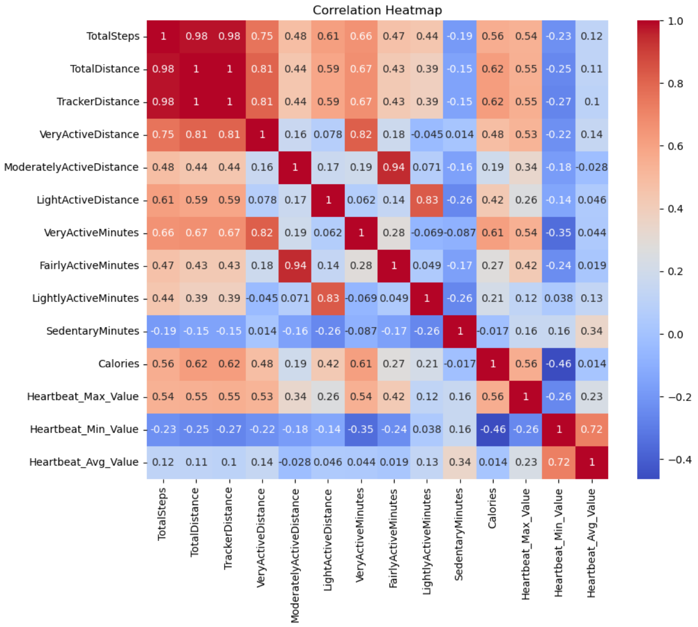


Figure 2: Correlation Heatmap

Insights:

Preliminary trends indicated a positive relationship between physical activity and sleep quality, laying the groundwork for clustering and predictive modeling.

6.3.2 Results

- A high-quality, unified dataset was created, ready for modeling.
- Missing data was effectively handled, reducing biases and preserving key patterns.
- Insights from EDA highlighted the interdependence of activity, sleep, and heart rate metrics

6.4 Phase 3: Predictive Modeling

The goal of this phase was to develop and evaluate predictive models that could reliably forecast sleep outcomes. This aligns with the overarching hypothesis that daily physical activity impacts sleep quality and quantity.

  - like Total Sleep Records were easier to predict, as the models showed consistent performance across training and test datasets.
  - Variability in Total Minutes Asleep and Total Time in Bed made them harder to predict, potentially due to unmeasured factors affecting sleep.

6.4.1 Dataset Preparation:

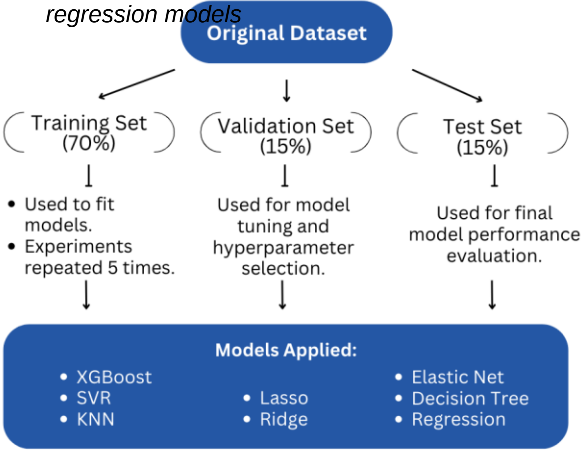
- **Features**: Key activity and physiological metrics (Total Steps, Calories, SedentaryMinutes, HeartRate_Avg, etc.).

- **Target Variables**:

    - Total Sleep Records
    - Total Minutes Asleep
    - Total Time in Bed

- **Splitting**:
    - The dataset was split into:
        - **Training Set (70%)**: To fit models.
        - **Validation Set (15%)**: For model tuning and hyperparameter selection.
    - **Test Set (15%)**: For final model evaluation. Splitting was repeated 5 times for robustness, and results were averaged.

Figure 6: Figure 6: Dataset preparation and splitting methodology, including applied regression models



- **Baseline Calculation**:
    - A baseline **Mean Squared Error (MSE)** was calculated for each target variable by predicting the mean of the training data for all samples in the test set.
    - This baseline serves as a benchmark to evaluate the effectiveness of machine learning models in reducing error and improving prediction accuracy.

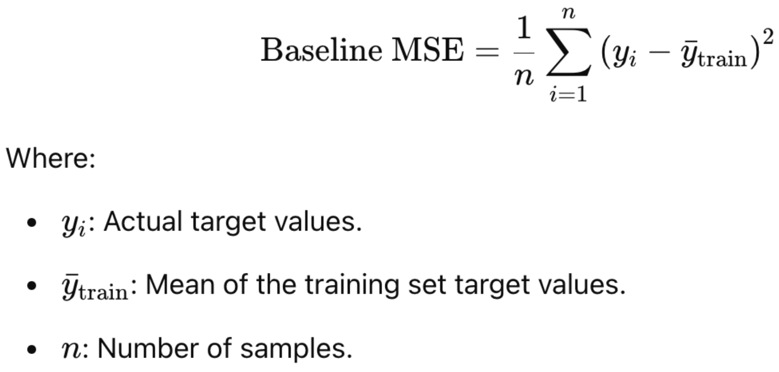
6.4.2 Model Selection:

- Multiple regression models were applied:
    - **Linear Regression**
    - **Ridge Regression**
    - **Lasso Regression**
    - **ElasticNet**
    - **Support Vector Regression (SVR)**
    - **XGBoost**
- Hyperparameter tuning was performed using GridSearchCV for all models.

6.4.3 Results

Baseline Performance

The **Baseline Mean Squared Error (MSE)** provides a simple yet meaningful benchmark to evaluate the performance of machine learning models. The baseline was calculated using the following formula, where predictions are made using the mean value of the target variable across the training dataset:



$$\text{Baseline MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}_{\text{train}}\right)^2$$

Where:

- y_i: Actual target values.
- $\bar{y}_{\text{train}}$: Mean of the training set target values.
- n: Number of samples.

Baseline MSE Results:

- **Total Sleep Records**: **0.0084**
- **Total Minutes Asleep**: **0.0355**
- **Total Time in Bed**: **0.0288**

Insights:

- The baseline serves as a benchmark for assessing the added value of predictive models.
- Models like XGBoost demonstrated significant improvements, reducing errors by over 70% compared to the baseline, underscoring their effectiveness.

Model Performance

The predictive modeling phase applied multiple regression algorithms to forecast sleep metrics. Key performance results are summarized below:

Total Sleep Records:

- **Best Model**: XGBoost

Model	Target	Average_Train_MSE	Average_Val_MSE	Average_Test_MSE
XGB	**TotalSleepRecords**	**0.0706**	**0.0858**	**0.1016**
SVR	TotalSleepRecords	0.1066	0.0893	0.104
KNN	TotalSleepRecords	0	0.0889	0.1006
Elastic Net	TotalSleepRecords	0.1028	0.089	0.1006
Lasso	TotalSleepRecords	0.1047	0.0896	0.1016
Linear	TotalSleepRecords	0.0982	0.0904	0.1006
Ridge	TotalSleepRecords	0.1001	0.089	0.1

Table 1: Performance of machine learning models for predicting sleep metrics – Total Sleep Records

Insights:

- XGBoost achieved the best performance, with the lowest validation and test errors.
- Predicting Total Sleep Records was relatively easier, as evidenced by consistent performance across datasets.

Total Minutes Asleep:

- **Best Model**: XGBoost

Model	Target	Average_Train_MSE	Average_Val_MSE	Average_Test_MSE
XGB	**TotalMinutesAsleep**	**0.0194**	**0.1095**	**0.1321**
SVR	TotalMinutesAsleep	0.1026	0.1198	0.1403
KNN	TotalMinutesAsleep	0	0.1206	0.1397
Elastic Net	TotalMinutesAsleep	0.1416	0.1542	0.1655
Lasso	TotalMinutesAsleep	0.1612	0.1774	0.1858
Linear	TotalMinutesAsleep	0.125	0.1354	0.1518
Ridge	TotalMinutesAsleep	0.1257	0.1364	0.1535

Table 2: Performance of machine learning models for predicting sleep metrics – Total Minutes Asleep

Insights:

- While XGBoost achieved the lowest error, overfitting was observed due to a significant gap between train and validation/test errors.
- The variability in this metric may arise from unmeasured factors affecting sleep duration.

Total Time in Bed:

- **Best Model**: XGBoost

Model	Target	Average_Train_MSE	Average_Val_MSE	Average_Test_MSE
XGB	**TotalTimeInBed**	**0.0168**	**0.1004**	**0.1174**
SVR	TotalTimeInBed	0.0916	0.1079	0.1294
KNN	TotalTimeInBed	0	0.1079	0.1255
Elastic Net	TotalTimeInBed	0.1243	0.1361	0.1475
Lasso	TotalTimeInBed	0.1433	0.1592	0.1655
Linear	TotalTimeInBed	0.1092	0.1172	0.1325
Ridge	TotalTimeInBed	0.1097	0.1181	0.1341

Table 3: Performance of machine learning models for predicting sleep metrics – Total Time in Bed

Insights:

- XGBoost provided the most accurate predictions, further reinforcing its suitability for complex, non-linear relationships in sleep data.
- Despite its strong performance, slight overfitting was observed, especially in metrics like Total Time in Bed.

7. Discussion

The findings of this project provide valuable insights into the relationship between physical activity and sleep outcomes, aligning with the hypothesis that daily activity impacts sleep quality and quantity. The project was structured into three phases—Data Cleaning and Preprocessing, Clustering Analysis, and Predictive Modeling—each contributing to a deeper understanding of the data and its implications.

7.1 Clustering Analysis

This phase aimed to identify behavioral patterns by grouping users based on their activity and sleep metrics. The goal was to understand how clustering could reveal distinct user groups and their associated sleep behaviors.

- **Key Results**:
    - The clustering analysis achieved a Silhouette Score of 0.33 after feature selection, indicating well-defined clusters.
    - **Cluster 0 (Active Users)**: Represented individuals with higher activity levels and better sleep outcomes, such as restful sleep and shorter times in bed.
    - **Cluster 1 (Sedentary Users)**: Represented individuals with high sedentary minutes and poorer sleep outcomes.
    - Statistical validation through ANOVA confirmed significant differences between clusters, with p-values < 0.05.

- **Implications**:
    - The clustering phase reinforced the hypothesis by linking active behaviors to better sleep and sedentary behaviors to poorer sleep.

    - These results offer opportunities for tailored health interventions, such as targeting sedentary users with activity recommendations to improve sleep outcomes.

7.2 Data Cleaning and Preprocessing

The first phase focused on ensuring the dataset's integrity and preparing it for analysis. Missing data was addressed using imputation techniques, and exploratory data analysis (EDA) provided critical insights into the relationships between variables.

- **Key Results**:
    - Positive correlations between activity metrics (e.g., Total Steps) and sleep outcomes (e.g., Total Minutes Asleep) supported the hypothesis.
    - Variability in sedentary minutes and heart rate metrics highlighted the diverse behavioral patterns in the dataset.

- **Implications**:
    - This phase ensured the dataset was ready for robust analysis, enabling reliable results in subsequent phases.
    - The initial correlations hinted at the potential for segmentation and prediction, guiding the objectives for clustering and modeling.

7.3 Predictive Modeling

The final phase sought to forecast sleep metrics using machine learning models. This phase demonstrated the ability to predict sleep outcomes and evaluate the effectiveness of different regression models.

- **Key Results**:

    - **Baseline Comparison**: Baseline MSE values were 0.0084 (Total Sleep Records), 0.0355 (Total Minutes Asleep), and 0.0288 (Total Time in Bed).
    - **Best Model**: XGBoost consistently outperformed other models, reducing errors by over 70% compared to the baseline.
        - **Total Sleep Records**: Test MSE of 0.1016, with stable performance across datasets.
        - **Total Minutes Asleep**: Test MSE of 0.1321, with variability due to external factors affecting sleep.
        - **Total Time in Bed**: Test MSE of 0.1174, highlighting XGBoost’s ability to handle complex relationships.
    - Challenges: Slight overfitting was observed for metrics like Total Minutes Asleep, indicating a need for additional features to improve generalization.

- **Implications**:
    - Predictive modeling validated the hypothesis by showing that activity metrics are strong predictors of sleep outcomes.

    - The results highlight the potential of machine learning in wearable health applications, enabling personalized recommendations for users.

7.4 General Implications

The results from all three phases demonstrate the significant impact of physical activity on sleep outcomes. The clustering analysis provided a segmentation framework that could inform personalized interventions, while predictive modeling showcased the potential for real-time forecasting and feedback.

Broader Impact

- **Health Recommendations**: The findings can inform the design of wearable devices and health applications to deliver actionable recommendations for improving sleep quality.
- **Future Research**: Incorporating additional variables, such as stress levels or dietary habits, could further enhance the understanding of activity-sleep interactions.
- **Real-World Applications**: These insights could be applied in wellness programs, encouraging users to adopt active lifestyles to achieve better sleep outcomes.

8. Contributions & Conclusions

This project explored the relationship between daily physical activity and sleep outcomes, contributing valuable insights into how behavioral patterns impact sleep quality and quantity. The results achieved across the three phases highlight the project's significance and provide actionable knowledge for the domain of health monitoring using wearable devices.

8.1 Contributions

- **Data Preparation and Quality Assurance**:
    - Developed robust preprocessing techniques, including imputation methods, to handle missing data effectively.
    - Performed exploratory data analysis to uncover key relationships between activity and sleep metrics, laying a solid foundation for clustering and predictive modeling.
- **Behavioral Insights through Clustering**:
    - Identified two distinct user groups:
        - **Active Users**: Exhibited better sleep quality and consistency, aligning with the hypothesis that higher activity improves sleep outcomes.
        - **Sedentary Users**: Demonstrated poorer sleep quality, characterized by longer durations and disrupted patterns.
    - Validated clustering quality with statistical methods, such as ANOVA, ensuring the reliability of identified patterns.

- **Predictive Modeling for Sleep Metrics**:
  - Developed machine learning models, with XGBoost consistently outperforming others by reducing prediction errors by over 70%.
  - Highlighted the predictive power of activity-related metrics (e.g., Total Steps, Sedentary Minutes) in forecasting sleep outcomes, such as Total Sleep Records and Total Time in Bed.
- **Actionable Insights for Personalized Health Monitoring**:
  - Provided a framework for using wearable device data to offer personalized recommendations for improving sleep quality.
  - Demonstrated the potential of machine learning in enhancing the utility of wearable health devices.

8.2 Conclusions

- **Validation of the Hypothesis**:
  - The project confirmed the hypothesis that daily physical activity significantly impacts sleep outcomes. Higher activity levels were associated with improved sleep quality, while sedentary behaviors correlated with poorer sleep patterns.
  - Additional, group associated with sedentary tendencies spent more time in bed and sleeping than the active group, but this does not mean better quality of sleep.

- **Significance of Clustering and Predictive Modeling**:
  - Clustering revealed meaningful behavioral groups, offering opportunities for targeted interventions to improve sleep outcomes.
  - Predictive modeling demonstrated that sleep metrics can be forecasted with reasonable accuracy, providing a foundation for real-time recommendations.

- **Practical Implications**:
  - These findings can inform the development of smarter wearable devices and health apps that deliver actionable feedback to users.
  - The framework established in this project could be expanded with additional features, such as stress levels or environmental data, to enhance predictive power.

- **Future Directions**:
  - Incorporating unmeasured factors (e.g., dietary habits, mental health) could improve the predictive accuracy of models.
  - Applying advanced machine learning techniques, such as deep learning, may yield further insights and enhance model performance.

---

9. References

1. Hiilloskorpi, H. K., Fogelholm, M., Laukkanen, R. M., Pasanen, M. E., & Oja, P. (2003). Estimation of energy expenditure using a heart rate monitoring method: Comparisons with indirect calorimetry. *Scandinavian Journal of Medicine & Science in Sports, 13*(3), 245–252. https://doi.org/10.1034/j.1600-0838.2003.00315.x

2. McClain, J. J., Lewin, D. S., Laposky, A. D., Kahle, L., & Berrigan, D. (2014). Associations between physical activity, sedentary time, sleep duration, and daytime sleepiness in U.S. adults. *Preventive Medicine, 66*, 68-73. https://doi.org/10.1016/j.ypmed.2014.06.003

3. Liao, Y., Robertson, M. C., Winne, A., Wu, I. H. C., Le, T. A., Balachandran, D. D., & Basen-Engquist, K. M. (2021). Investigating the within-person relationships between activity levels and sleep duration using Fitbit data. *Translational Behavioral Medicine, 11*(2), 619–624. https://doi.org/10.1093/tbm/ibaa071

4. Nakanishi, M., Izumi, S., Nagayoshi, S., Kawaguchi, H., Yoshimoto, M., Ando, T., Nakae, S., Usui, C., Aoyama, T., & Tanaka, S. (2018). Estimating metabolic equivalents for activities in daily life using acceleration and heart rate in wearable devices. *BioMedical Engineering Online, 17*(1), 100. https://doi.org/10.1186/s12938-018-0532-2

5. Park, H., Noh, J., & Lee, J. (2024). Wearable-based prediction of sleep outcomes using physical activity, light exposure, and heart rate variability: A machine learning approach. *International Journal of Artificial Neural Networks*, 56(2), 240-251. https://doi.org/10.1016/j.iann.2024.2405077

6. Karimi, F., Amoozgar, Z., Reiazi, R., Hosseinzadeh, M., & Rawassizadeh, R. (2024). Longitudinal analysis of heart rate and physical activity collected from smartwatches. *CCF Transactions on Pervasive Computing and Interaction, 6*(1), 18–35. https://doi.org/10.1007/s42486-024-00147-y

7. Nauha, L., Farrahi, V., Jurvelin, H., Jämsä, T., Niemelä, M., Ala-Mursula, L., Kangas, M., & Korpelainen, R. (2024). Regularity of bedtime, wake-up time, and time in bed in mid-life: Associations with cardiometabolic health markers with adjustment for physical activity and sedentary time. *Journal of Activity, Sedentary and Sleep Behaviors, 3*(2). https://doi.org/10.1186/s44167-023-00040-6