# Final Report: Special Topics in Analytics

By: Priyanka Bhosale
Date: July 25, 2024

## Introduction

This report delves into the analysis and interpretation of workout data from patients of MyoCycle, a specialized exercise device designed to help individuals who have lost the ability to work out independently. The analysis focuses on identifying key features that impact patient longevity and consistency in their workout routines. This report is structured to provide definitions, methodological approaches, questions addressed, and the conclusions drawn from the analysis.

### Summary Data Features

The summary data includes the following features:

- user_id: Unique identifier for each patient
- device: Type of device used
- duration_min: Duration of the workout in minutes
- avg_power_mi: Average power generated per mile
- date_time: Timestamp of the workout session

### Engineered Features

To measure longevity and consistency, several engineered features were created:

1. total_workout_duration: Total workout duration for each patient.

2. workout_consistency_sd: Standard deviation of the days between workouts.
   This feature was engineered to measure the variability in the **number of days between workouts** for each patient
   Calculate the difference in "days_between" each consecutive workout.
   Calculate the Standard Deviation of the "days_between" for each patient

3. sustained_engagement: Time span from the first to the last workout for each patient.

4. duration_min_sd: Standard deviation of workout durations, indicating consistency and fatigue.
   This feature was engineered to measure the variability in the **duration of workouts.**
   A low standard deviation means the durations are consistent and implies less fatigue

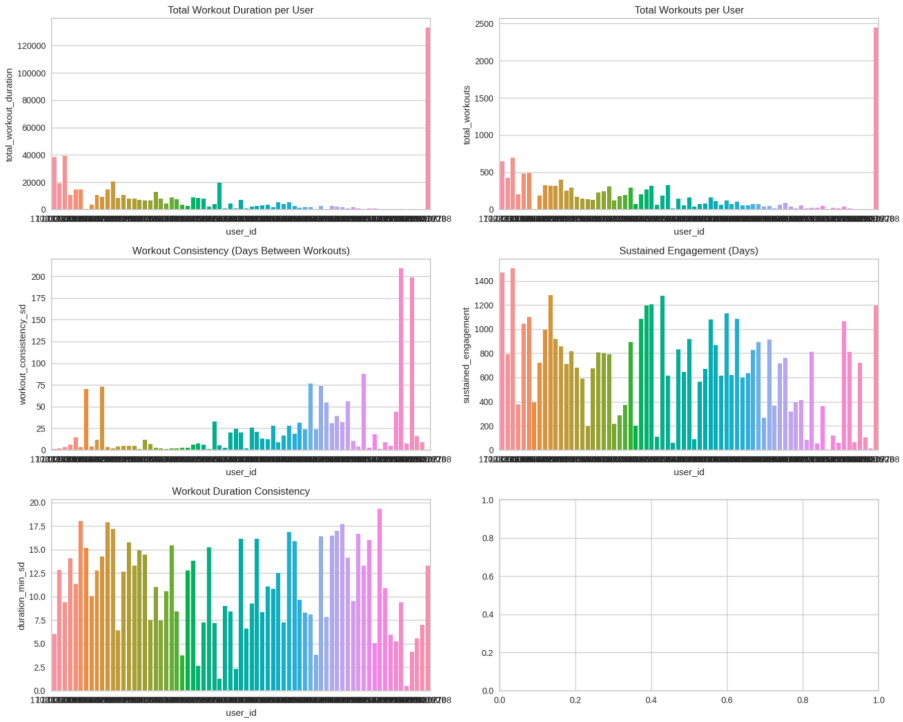
5. weekly_avg_power: Average power output per week.

6. high_consistency (for approach-1): Binary indicator of workout consistency, based on a threshold of workout consistency (standard deviation).
   - Set it to 0 (No) if the standard deviation of the workout consistency for the patient > threshold (median standard deviation of workout consistency)
   - Set is as 1 (Yes) if the standard deviation of the workout consistency for the user < threshold (median standard deviation of workout consistency)

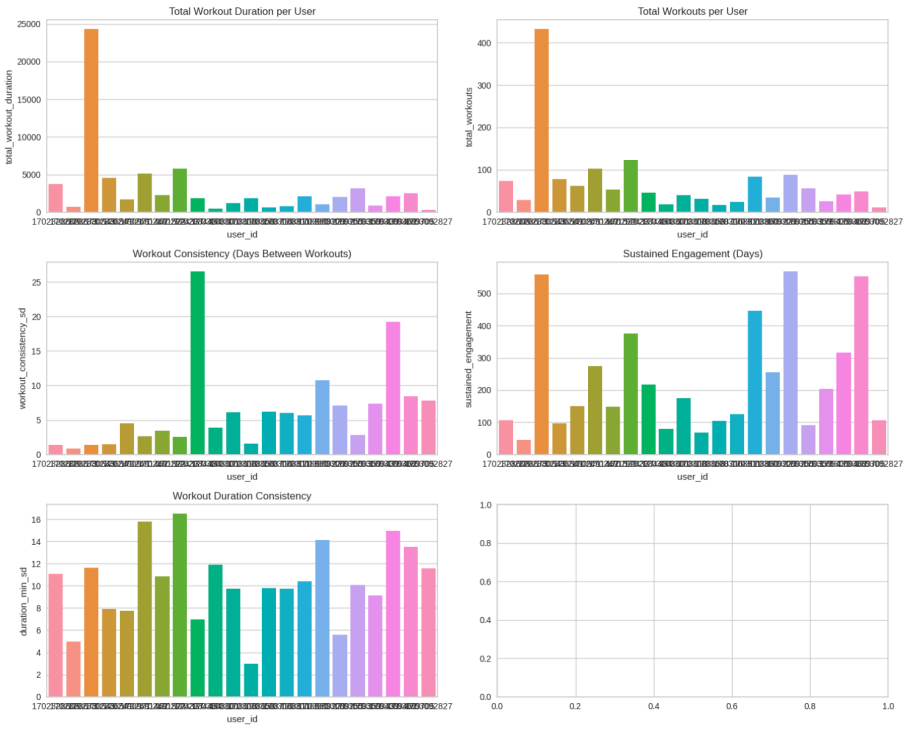
7. high_consistency (for approach-2): Binary indicator of consistency, based on a threshold of workout duration standard deviation.
   - Set it to 0 (No) if the standard deviation of the workout duration for the user > threshold (median standard deviation of workout duration)
   - Set is as 1 (Yes) if the standard deviation of the workout duration for the user < threshold (median standard deviation of workout duration)

## Methodological Approaches

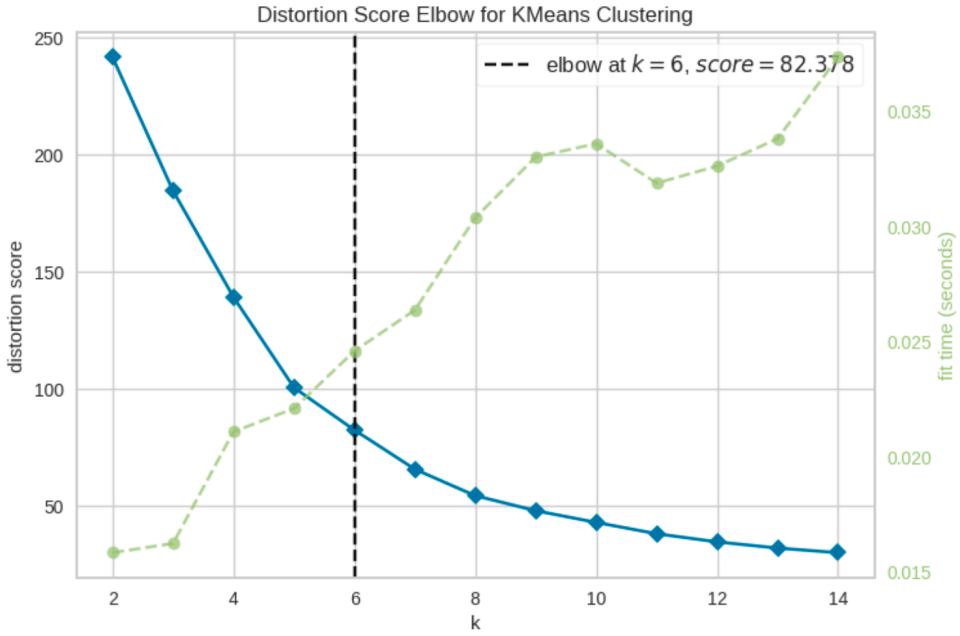
**1.** Clustering (KMeans):
Clustering was performed using KMeans on the longevity features.
The dataset was split into a training set (80%) and a test set (20%).
- Device-1 dataset: 6 clusters (0 to 5)
- Device-2 dataset: 7 clusters (0 to 6)



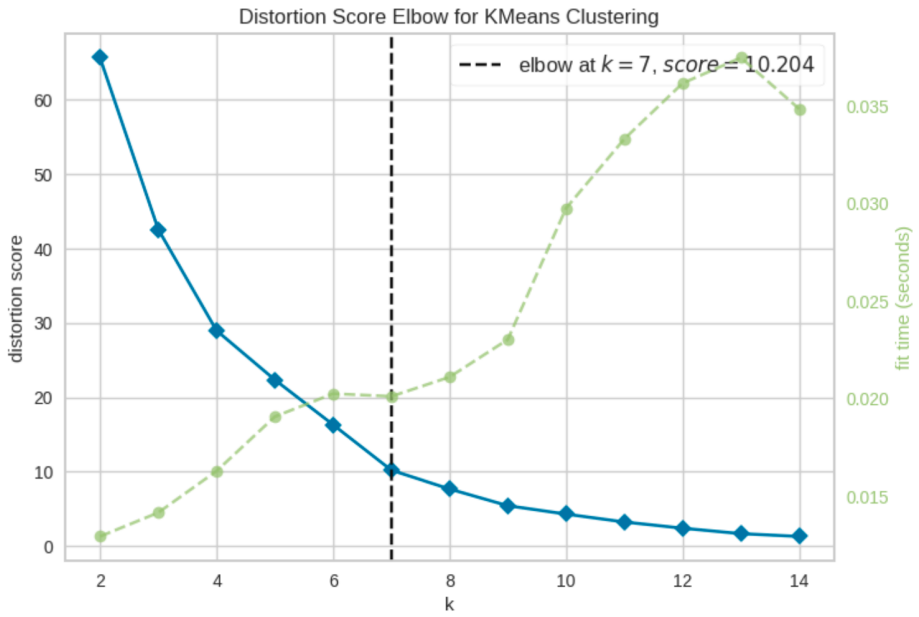

Device-1 patient data with engineered features: Total Patients 71




Device-2 patient data with engineered features: Total Patients 22




Optimal number of clusters/patient subtypes for Device-1 Dataset using Elbow Method
Total Clusters: 6 (0 to 5)




Optimal number of clusters/patient subtypes for Device-2 Dataset: Elbow Method
Total Clusters: 7 (0 to 6)

2. **Hypothesis Testing:** Hypothesis testing was conducted using t-tests to determine the significance of different features.

3. **Random Forest Classifier:** A Random Forest Classifier was used to identify the feature importance for both Device-1 and Device-2 datasets.

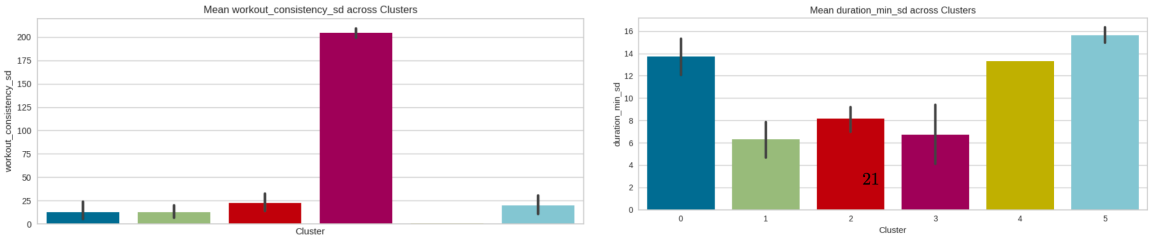
## Questions and Analysis

**Question 1:** Optimal Patient Subtypes (Clusters) for Longevity (Consistency)

The clusters were analyzed to determine the best to worst consistency order for both Device-1 and Device-2 patients.

Device-1 Consistency Cluster Order (best to worst):

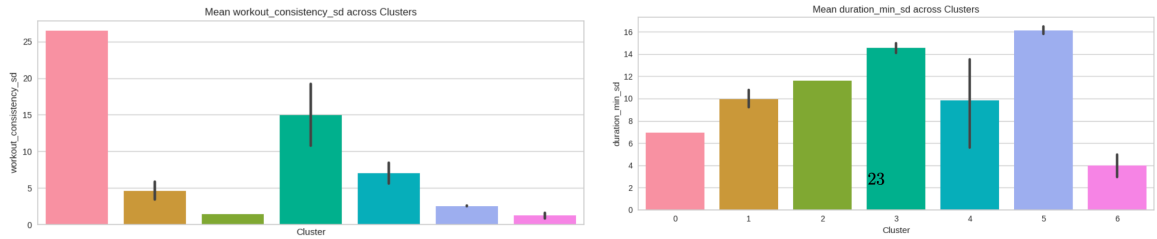
1. Cluster 4
2. Cluster 1
3. Cluster 0
4. Cluster 5
5. Cluster 2
6. Cluster 3




Mean calculation for workout_consistency_sd and duration_min_sd across patient subtypes of Device-1

Device-2 Consistency Cluster Order (best to worst):

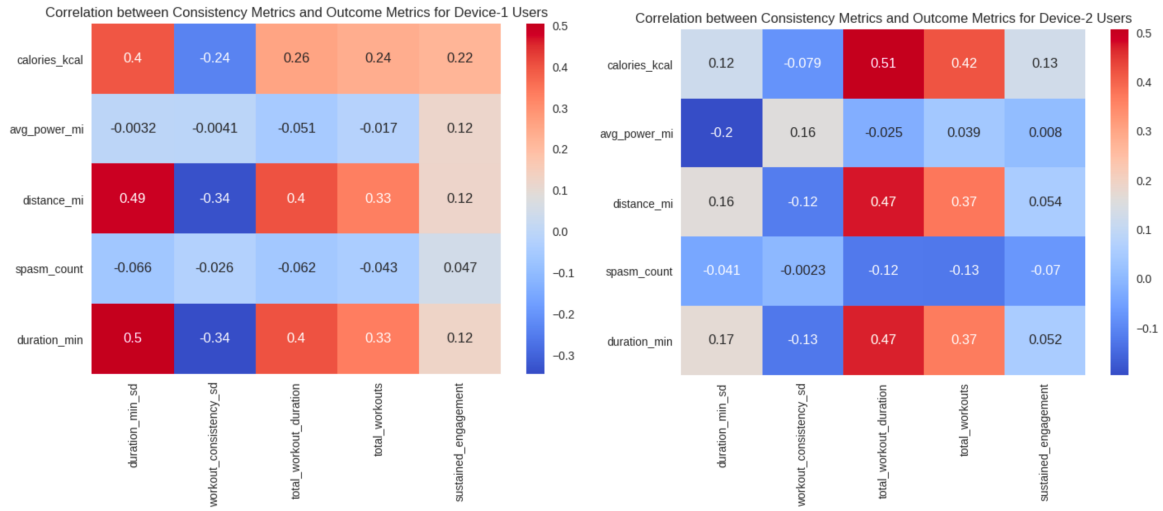
1. Cluster 6
2. Cluster 2
3. Cluster 5
4. Cluster 1
5. Cluster 4
6. Cluster 3
7. Cluster 0




Mean calculation for workout_consistency_sd and duration_min_sd across patient subtypes of Device-2

**Question 2:** Impact of Spasm Count on Longevity (Consistency)

The correlation between spasm count and longevity metrics (Calories, Distance, Duration) was analyzed. It was found that none of the longevity metrics had a high correlation with spasm count.




Correlation Metrics for Device-1 and Device-2 patients

**High Correlations for Device-1 and Device-2 patients:**

- Calories: Workout Duration SD, Total Workout Duration, Total Workouts, Sustained Engagement
- Distance: Workout Duration SD, Total Workout Duration, Total Workouts, Sustained Engagement
- Duration: Workout Duration SD, Total Workout Duration, Total Workouts, Sustained Engagement

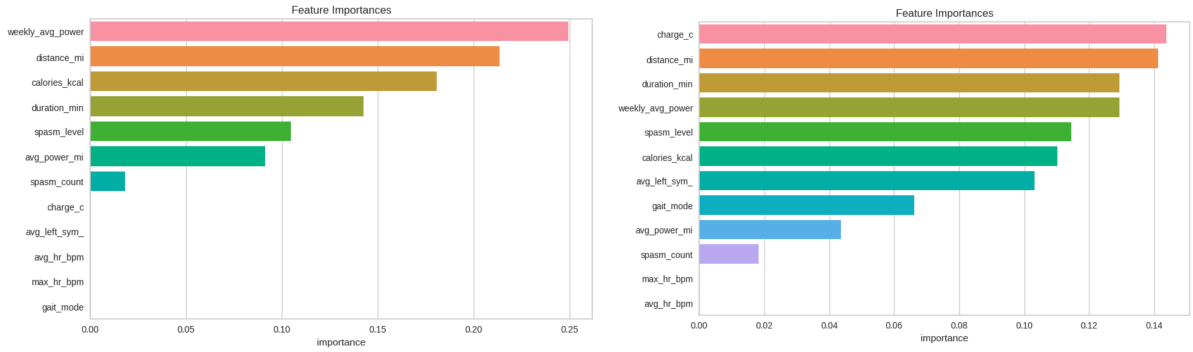
## Approach

**Approach 1:** Significant Features for High Consistency

The median consistency (SD) was used as the threshold to classify patients into high and low consistency groups.

Definitions:

- High Consistency: Patients with consistency (SD) greater than the threshold.
- Low Consistency: Patient with consistency (SD) less than the threshold.




Device-1 vs Device-2 Feature Importance

Top 4 Important Features:

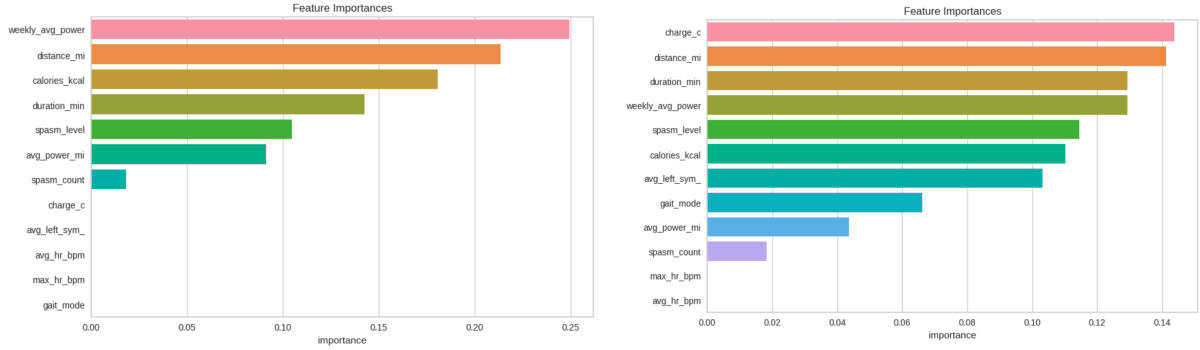
- Device-1: Weekly Average Power, Distance, Calories, Workout Duration.
- Device-2: Charge, Distance, Duration, Weekly Average Power.

**Approach 2:** Significant Features for High Consistency

A similar method was applied but, in this case, using the median workout duration (SD) as the threshold.

**Conclusion for Approach-1 and Approach-2:**

The top 4 important features identified in both approaches matched 100%.




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| False | 0.89 | 0.93 | 0.91 | 1352 |
| True | 0.92 | 0.88 | 0.90 | 1268 |
| accuracy | | | 0.91 | 2620 |
| macro avg | 0.91 | 0.91 | 0.91 | 2620 |
| weighted avg | 0.91 | 0.91 | 0.91 | 2620 |

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| False | 0.94 | 0.98 | 0.96 | 177 |
| True | 0.97 | 0.91 | 0.94 | 128 |
| accuracy | | | 0.95 | 305 |
| macro avg | 0.96 | 0.95 | 0.95 | 305 |
| weighted avg | 0.96 | 0.95 | 0.95 | 305 |

Device-1 vs Device-2 Random Forest Classifier Results

## Conclusions

Motivating Factors for Patients

The analysis identified key motivating factors for patients using both Device-1 and Device-2.

Device-1:

- Weekly Average Power
- Distance
- Calories
- Workout Duration

Device-2:

- Charge
- Distance
- Duration
- Weekly Average Power (converts to efficiency)

## Final Insights

Consistency in maintaining weekly average power is a likely factor for patients to continue their workout routines. The identified features can help customize workout plans and improve patient engagement and longevity with the MyoCycle devices.