Fitbit Data Analysis

Machine Learning
Group 8

Julia Ciriello
Nik Gupta
Devin Ridley
Alvaro Rivera-Eraso

Table of Contents

Objective

The goal of this analysis was to explore the various ways in which health tracker data could be used to uncover meaningful insights connecting a person's level of activity and sleep, among other factors. We approached the data with a scientific mindset rather than seeking to produce a meaningful business decision.

We chose the following objectives and initial hypotheses:

1. *Cluster observations into meaningful groupings, to associate sleep, exercise, and activity.* We predicted that there would be natural clustering of observations of sleep, exercise, and activity. Given the nature of the data, we set out to understand the participants and to extract meaningful insight from the patterns within their usage. This insight could be useful to both users or government health agencies in order to better understand habits that lead to better health, including sleep and exercise.
2. *Predict the quality of a person's sleep based on the amount and quality of their activity throughout the day.* We predicted that higher quality of sleep will follow higher duration and intensity of exercise. Similarly, the output of the classification model could be used to provide better health guidance and promote healthy sleep and exercise habits.

Data Preparation

Data Description

Source

The data for our analysis was drawn from https://www.kaggle.com/datasets/arashnic/fitbit, a Kaggle dataset with high quality, high resolution data from 30 volunteer participants who consented to provide data from their Fitbit wearable trackers. The dataset was rated a 10 for usability by the Kaggle community. We downloaded the data and saved it to a shared folder, from which we collaborated.

Data Overview

This data represents the activity monitoring from 30 fitbit users who allowed their data to be distributed. These trackers monitored the vitals of the participants as well as their activity, including sleep tracking. The data was collected at various frequencies, with minute level being the most frequent.

Exploratory Data Analysis

To begin our analysis, we started with a comprehensive overview of the data to understand its structure and any underlying issues that would interfere with our analysis.

Format of Data

The first challenge with our data was that it was provided in an array of files rather than a single flat file. The directory contains 18 files, which are named to describe the type of data and the data's level of detail.

Table 1: Summary of files provided

	Level of Aggregation			
Data Type	Day	Hour	Minute	Second
Sleep	sleepDay_marged	*N/A*	minuteSleep_merged	*N/A*
Calories	dailyCalories_merged	hourlyCalories_merged	minuteCaloriesNarrow_merged minuteCaloriesWide_merged	*N/A*
Exercise	dailyIntensities_marged	hourlyIntensities_merged	minuteIntensitiesNarrow_merged minuteIntensitiesWide_merged	heartrate_seconds_merged
Steps	dailySteps_merged	hourlySteps_merged	minuteStepsNarrow_merged minuteStepsWide_merged	*N/A*
METs	*N/A*	*N/A*	*N/A*	minuteMETsNarrow_merged

In addition, there was one file (weightLogInfo_merged) that reported weight and did not have a defined interval. There was also a pre-aggregated daily file named dailyActivity_merged. The 'Wide' and 'Narrow' files noted in the table above differed in that the wide files used a separate column per minute and the narrow used a separate record per minute.

Removing Unnecessary Data

We opted not to use METs data since Fitbit did not disclose their algorithm for calculating METs and we determined that it was not necessary to include it given the other data available to capture activity intensity. https://community.fitbit.com/t5/Get-Moving/METs-minutes-vs-Active-Minutes/td-p/1503733

Data Engineering

We decided against starting with the pre-aggregated dailyActivity_merged dataset because it only had 940 records and did not contain the sleep data that we wanted to analyze. We created our own dataset at the hourly level using the below files:

Table 2: Files used and data preparation steps

Data Type	File name	Data Preparation Steps
Steps	hourlySteps_merged	No additional preparation needed, aside from converting the ActivityHour field to Pandas' datetime format. 22,099 records.
Calories	hourlyCalories_merged	No additional preparation needed, aside from converting the ActivityHour field to Pandas' datetime format. 22,099 records.

Sleep	minuteSleep_merged	File contains sleep values summarized for each user by minute. Sleep values: 1 = Asleep, 2 = Restless, 3 = Awake The first step was to create 3 new columns (one for each sleep value) to have separate counts by sleep type. Once this was complete we resampled the data to summarize the new columns by hour instead of by minute. Result: 3,574 records
Exercise	minuteIntensitiesWide_merged	File contains intensity values for each user by minute (60 columns each representing 1 minute). Intensity values: 0 = Sedentary, 1 = Light, 2 = Moderate, 3 = Very Active The first step was to create 4 new columns (one for each intensity value) to have separate counts by intensity type. Once this was complete we resampled the data to summarize the new columns by hour instead of by minute. Result: 21,645 records

With the above 4 datasets prepared, we used Pandas' merge() function to combine them and provide us with the following final data for use in our modeling:

Table 3: Summary of final dataset for modeling

DatetimeIndex (ActivityHour): 22458 entries, 2016-04-12 00:00:00 to 2016-04-11 23:00:00			
#	Column Name	Non-Null Count	Dtype
0	Id	22458	object
1	StepTotal	22099	int64
2	Calories	22099	int64
3	0=sedentary	21645	int64
4	1=light	21645	int64
5	2=moderate	21645	int64
6	3=veryactive	21645	int64
7	1=asleep	3574	int64
8	2=restless	3574	int64
9	3=awake	3574	int64

Missing Data

Sleep data was only available for around 16% of the total records. We expected around double this availability (i.e. ~33%, assuming 8 hours of sleep per day), however sleep data was only able to be automatically recorded for a limited range of Fitbit models, while other users had to enter their data manually. As such, we were limited to 3,574 records containing sleep data. The data was also missing a small amount (~2%) of intensity data. We used the fillna() function to populate all missing records with 0.

Model Design

Choice of Models

In order to derive meaningful insight from the data, we chose to create both clustering models and classification models. For each, we tried different model types and determined the best single approach for this dataset.

Clustering Model

For the clustering stage, two models have been considered: a KMeans model and a hierarchical clustering model. To begin this procedure, the selection of target variables was initiated. Consequently, the entire matrix of variables comprising the data frame was taken, with the exception of the subjects' IDs and the measurement times.

Model Preparation Steps

To prepare the data for the clustering model, we first created two new columns named 'SleepMin' and 'ActiveMin'. For ‘SleepMin’ we decided to only include the ‘1=asleep’ and ‘2=restless’ minutes, and ignored ‘3=awake’. Similarly for ‘ActiveMin’ we only included ‘1=light’, ‘2=moderate’, and ‘3=veryactive’, ignoring ‘0=sedentary’. The reason for excluding the awake and sedentary minutes was to find clusters relating to “true” sleep and exercise behaviors. We also normalized these two new fields to ensure they did not exceed 60 minutes, as there appeared to be a few erroneous records showing 61 minutes or more in an hour. The resulting dataframe, after dropping the columns not needed, had the following properties:

Table 4: Numeric column details

	StepTotal	Calories	SleepMin	ActiveMin
count	22458	22458	22458	22458
mean	315.048357	95.829994	8.281548	9.376881
std	686.019973	61.441794	20.039895	12.9495
min	0	0	0	0
25%	0	62	0	0
50%	33.5	83	0	3
75%	349	107	0	15
max	10554	948	60	60

After that stage was completed, the data frame was grouped by the 'Id' column, and the data was resampled at a daily frequency using the resample function with a frequency of '1D'. Through testing we noticed the ‘Id’ column skewed the clustering results, so we created a copy of our dataframe with 'Id' dropped, and then computed the first quartile (Q1), third quartile (Q3), and the interquartile range (IQR) values of the dataframe using the quantile function.. Next, the df_daily_avg_excSlp data frame was filtered to include only the rows whose values fell within the established limits, and missing values in the dataframe were filled with the column means using the fillna(X.mean()) method.

K-Means Clustering Stage

In order to determine the optimal number of clusters in the K-Means model, we first decided to take a look at the elbow method and silhouette method. The corresponding graphs for these methods can be found in Appendix 1.

Examination of both the elbow method and the silhouette analysis suggested that the appropriate number of clusters for this dataset is k=3. Regarding the inertia plot, there is a noticeable decrease in inertia as we move from k=2 to k=3 clusters. This observation indicates that utilizing 3 clusters better captures the inherent sparsity of the data. Analyzing the results from the silhouette analysis, the highest score is consistently associated with k=2 clusters, with a value of 0.554. This highlights that using 2 clusters achieves a good balance between identifying distinct patterns and maintaining internal cohesion within each cluster. However, taking into consideration the previous results, a silhouette score of 0.517 is also obtained with k=3 clusters. Careful observation of the elbow graph suggests that opting for 3 clusters results in a more substantial reduction in the overall inertia. Therefore, selecting 3 clusters is a preferable approach for the initial K-Means model.

Then, to validate our observations we performed a hyperparameter grid search by using the GridSearchCV function. The parameter grid was set to evaluate values ranging from 2 to 8 clusters. The dataset was divided into 5 folds for cross-validation during the grid search process. After completing the grid search, the best number of clusters was determined to be 8. This result does not align with the previous findings from the contour and elbow analysis; however, it highlights how differing approaches to tuning hyperparameters can result in very different results.

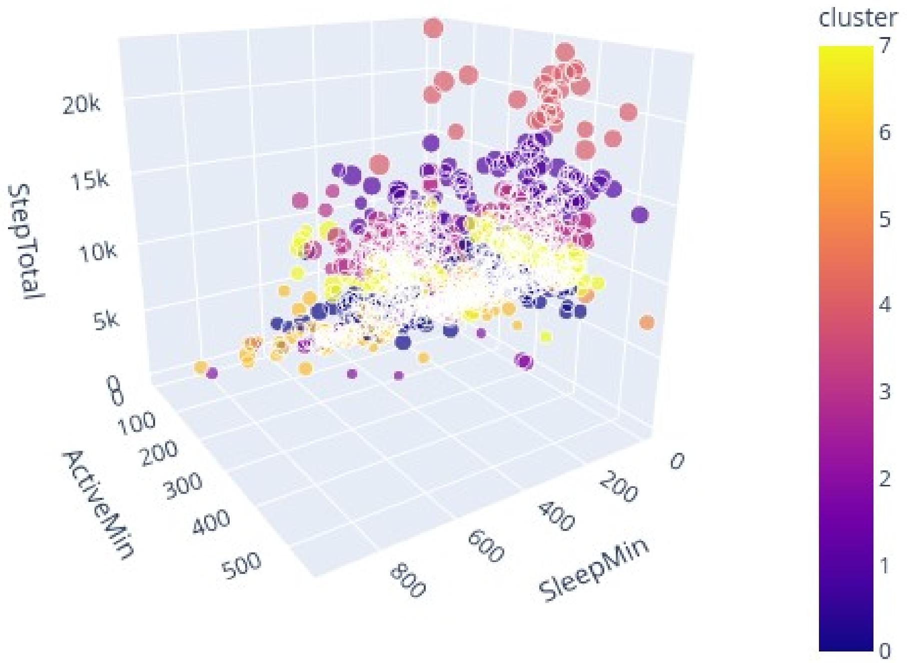
Subsequently, a K-means algorithm was instantiated with the optimal number of clusters (n_clusters=8) and fitted to the data set using the kmeans.fit(X) method. The below table summarizes the patterns noted for each of the clusters; more detailed description can be found in Appendix 2:

Table 5: Overview of K-Means clusters

Cluster #	Overview of observations
1	Over 185 observations, daily averages include StepTotal of 6,111 and Calories burned at 2,203. Sleep averages 188 mins, ActiveMinutes 244. Sleep and Active minutes seem inversely correlated.
2	Data from 106 observations reveals high activity (StepTotal: 14,311, ActiveMinutes: 328), but with SleepMinutes at 208, possibly indicating inconsistent sleep tracking.
3	Across 162 observations, around 283 steps and 1,816 calories daily, with about 59 mins of sleep and 24 active mins, suggesting restful routines.
4	Over 166 observations, active with 11,000 steps, 2,540 calories burned, and around 291 sleep mins. Median active mins: 293.
5	In 29 observations, high intensity (19,000 steps, 3,440 calories), with minimal sleep and consistent activity (333+ mins).
6	With 7,370 steps, 3,107 calories, this cluster's sleep tracking is absent (0 mins), activity spans 352 to 552 mins.
7	Across 151 observations, light activity (StepTotal: 3,311, ActiveMinutes: 148), varying sleep (231 mins), caloric intake 1,924.
8	Over 158 observations, highly active (StepTotal: 8,649, ActiveMinutes: 290), caloric intake 2,439, sleep averages 215 mins.

Throughout these clusters, the recurring presence of zero sleep duration instances draws attention to potential inaccuracies in sleep tracking. The varying relationships between sleep duration and activity levels underscore the complexity of balancing physical engagement and restful sleep. Further investigation into these relationships could shed light on individual preferences, lifestyle choices, and the impact of tracking errors on our understanding of sleep patterns.

Figure 1: K-means clustering 3D chart



***Note**: To see the interactive figure refer to the notebook called Part2_Clustering_3models.ipynb*

Hierarchical Clustering Stage

The hierarchical clustering procedure began with the creation of a dendrogram plot, which allows for visual observation of the group distribution along with the distances they encompass. This provided a general idea of the data distribution for each group, revealing the presence of primarily 2 clusters highlighted in blue. Subsequently, the existence of 4 major branches became apparent, with two in orange and two in green. The dendrogram corresponding to the aforementioned structure is included in Appendix 3.

As a second step in identifying the optimal number of clusters, silhouette analysis was conducted over various values of K = [2, 3, 4, 5, 6, 7, 8, 9]. This analysis revealed a significant drop in the silhouette score between cluster 2 and cluster 3. The silhouette graph is also presented in Appendix 3. Analyzing these silhouette scores, the most notable observation is that the highest value is attained when considering k=2, resulting in a silhouette score of approximately 0.52. This indicates that, according to the silhouette method, a partition into two clusters is deemed the most suitable configuration for this dataset. In the context of hierarchical clustering, the silhouette score captures how well the hierarchical structure separates the data points into distinct groups. Thus, in this specific case, a silhouette score of around 0.52 at k=2 implies that the two identified clusters are relatively well-defined and distinct from each other. The data points within each cluster are more similar to each other compared to the points in the other cluster.

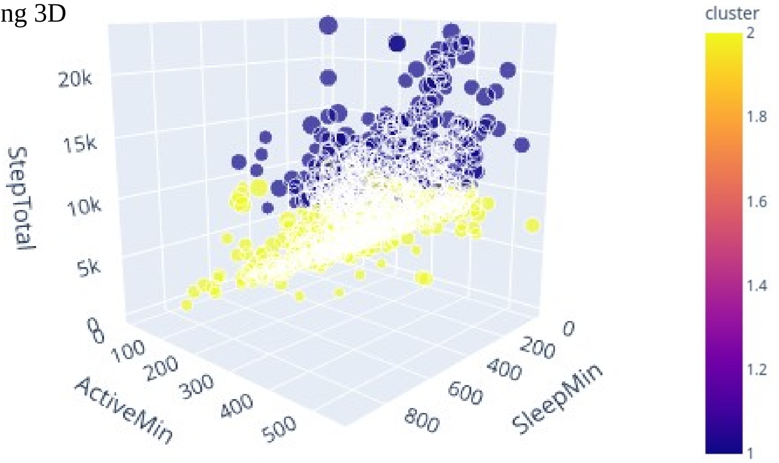
Now that we are aware that the most suitable number of clusters is k = 2, the next step involves allocating observations to the previously determined optimal clusters. This process is followed by conducting a descriptive analysis for each cluster. This practice aids in comprehending the distinctive characteristics and specific attributes of observations within each group identified through hierarchical clustering.

Table 6: Overview of Hierarchical Clustering clusters

Cluster #	Overview of observations
1	Cluster of 323 observations examines metrics: moderately active (mean StepTotal: 12,803, SD: 2,951), balanced energy intake (mean Caloric intake: 2,706). Sleep varies (mean SleepMin: 242 mins, SD: 222), from none to long hours (0 to 679 mins). ActiveMin mean: around 303 mins, showing significant physical activity.
2	The second cluster, with 637 observations, shows less activity (mean StepTotal: 4,478, SD: 3,091) and moderate energy intake (mean Caloric intake: 2,089). Sleep varies widely (mean SleepMin: 169 mins, SD: 233), including instances of no sleep. Physical activity is lower (mean ActiveMin: 177 mins) compared to Cluster 1.

In both clusters, it's important to note that the range of SleepMin is wide, indicating substantial variation in sleeping patterns. Factors such as lifestyle choices, health conditions, and individual preferences could play a role in these variations. Additionally, the distribution of sleep duration is skewed towards lower values in Cluster 2, potentially indicating a greater prevalence of individuals with shorter sleep durations associated with lower activity levels, or a mistake in the data capture. This analysis highlights the complex interplay of activity levels, caloric intake, and sleep duration within each cluster and underscores the need for further investigation into the underlying factors influencing these patterns.

Figure 2: Hierarchical clustering 3D



***Note**: To see the interactive figure refer to the notebook called Part2_Clustering_3models.ipynb*

Classification Model

The goal of the classification problem is to determine whether one can predict whether a person will get a good night's sleep by using their Fitbit activity data. Further data preparation was required before attempting the classification problem. Firstly, the measure "Total Sleep Time" was defined as the sum of minutes a person spends trying to sleep, either asleep, restless, or awake (as measured by their Fitbit). Many participants in the study did not wear their Fitbit to sleep overnight. Any participants who didn't reliably sleep with their Fitbit on (less than an average of 4 hours of Total Sleep Time) over the course of the month were removed from consideration. In order to analyze the activity level during the day with the corresponding night's sleep, the date cut-off time needed to be shifted. By looking at the distribution of Total Sleep Time, the day was defined as 08:00 to 07:59. For example, Monday's data should start on Monday at 08:00 and end on Tuesday at 07:59 in order to match the day's activity with a full night's sleep. After shifting the dates accordingly, the data was grouped by date and ID. Within the users who regularly wore their Fitbit to sleep, there were still individual nights where minimal sleep data was available. Records with less than 3 hours of Total Sleep Time were dropped. A correlation matrix of the data at this stage (Appendix 5) shows very weak correlations between variables, outside of the groups of dependent variables (i.e. sleep variables are related to one another; activity variables are related to one another).

To perform the classification, we defined "Good Sleep" as a boolean variable indicating whether someone slept at least 7 hours in a night (using the "asleep" field, not the Total Sleep Time). The independent variables in the model are StepTotal, Calories, and level of activity (sedentary, light, moderate, veryactive); the dependent variable is Good Sleep. The MinMaxScaler was applied and a train/test split stratified by class was created. As a baseline, the naive classifier (which classifies all data into the majority class) resulted in an F1 score of about 0.74 and an accuracy score of about 0.59 for both the training and testing sets.

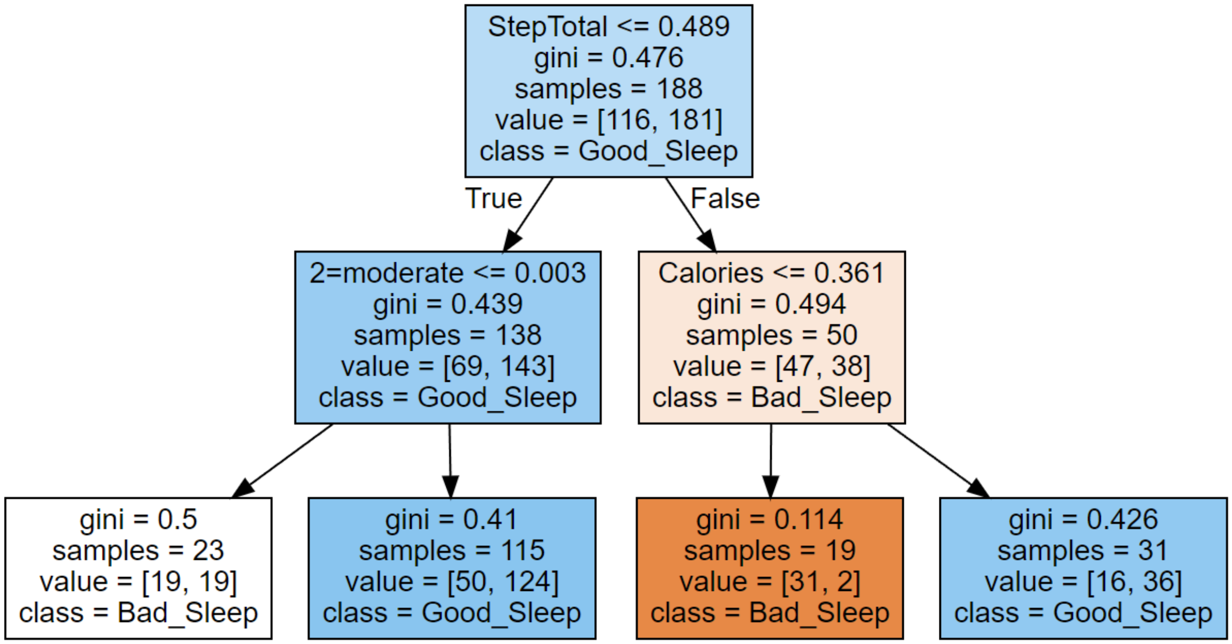
Classification Model Evaluation

By trying a wide array of classifiers and pairing that with hyperparameter tuning, we hope to be able to identify the best classifier for this data. The following classification models were created: Logistic Regression, Random Forest Classification, Gradient Boosting Classifier, AdaBoost Classifier, and Extra Trees Classifier. The first three models underwent hyperparameter tuning to improve model performance. The Logistic Regression Model was the only model to underperform the naive classifier on both F1 and accuracy scores. The Random Forest model resulted in the best F1 score (0.768) and the Extra Trees model resulted in the best accuracy score (0.697) which both constitute low to moderate improvements over the naive classifier.

Table 7: Comparison of various classification models' scores

Model	**Test F1 Score**	**Test Accuracy Score**
Naive Classifier	*0.739*	*0.586*
Logistic Regression	0.724	0.576
Random Forest Classification	**0.768**	0.677
Gradient Boosting Classifier	0.736	0.667
AdaBoost Classifier	0.715	0.646
Extra Trees Classifier	0.766	**0.697**

Figure 3: Visual of one tree in Random Forest Classification Model



The improvement in F1 score was very minimal across all models compared to the naive classifier which suggests that there is at best a tenuous connection between activity level during the day and a sleep of at least 7 hours. It is likely that activity level is only one small factor influencing sleep and a better model could be built if it included more data such as the participant's stress levels, work schedule, eating habits, alcohol/drug consumption, and family status (to name only a few).

Conclusions

Through the course of our analysis, we generated both clustering and classification models. Our conclusions can therefore be divided in the same way.

Clustering of Daily Data

The clustering models were able to effectively group daily Fitbit data by using a combination of factors from the data.

Regarding the K-means model analysis, a significant data distribution is noticeable, forming distinct and well-separated groups. However, when comparing the performance of both models, it's clear that the hierarchical clustering model took a more suitable approach in generating group distribution. Therefore, the formation of groups in this model displays a better equilibrium, achieving a favorable balance in the observations.

Although the assessment of sleep duration showed inadequacies and patterns of 0-minute sleep periods, the hierarchical model managed to distribute these sleep periods more effectively, resulting in well-balanced groups that reflect a more typical daily routine for multiple individuals.

In conclusion, while both models performed reasonably well, the K-means model demonstrated slightly inferior performance. Nevertheless, it's important to highlight that the data distribution into 8 clusters revealed a deficiency in measuring the sleep duration variable. For this reason, it is considered necessary and advisable to run more than one model in order to obtain different perspectives and points of view that make it possible to highlight flaws and important points in the dataframe.

Classification of Sleep Quality

The classification models showed limited improvement over the naive classifier. This held true for all five models that were attempted, which suggests that it is inherent in the data and/or problem rather than the particular models we used. The key challenges in creating the classification models were the limited time frame and number of participants, as well as a lack of other personal data which would better explain the participants’ quality of sleep.

Next Steps

In order to build off of our clustering model and hopefully build a classifier with good performance, we identified a number of improvements that would better allow us to meet our stated objectives and generate more impactful insights.

Participants and Time

One of the main limitations we found in our data set was that we lacked the quantity of data to build extremely robust models. If we were to have more participants and track data over a longer time frame it’s very likely that our classification models in particular would be improved.

Weight Data

While weight data was included in the data set, it was sparse and lacked consistency in its collection time. Knowing that weight is an important feature and target for health-related modeling, having a daily interval and consistency in the weight data would be very helpful so we could use it for modeling.

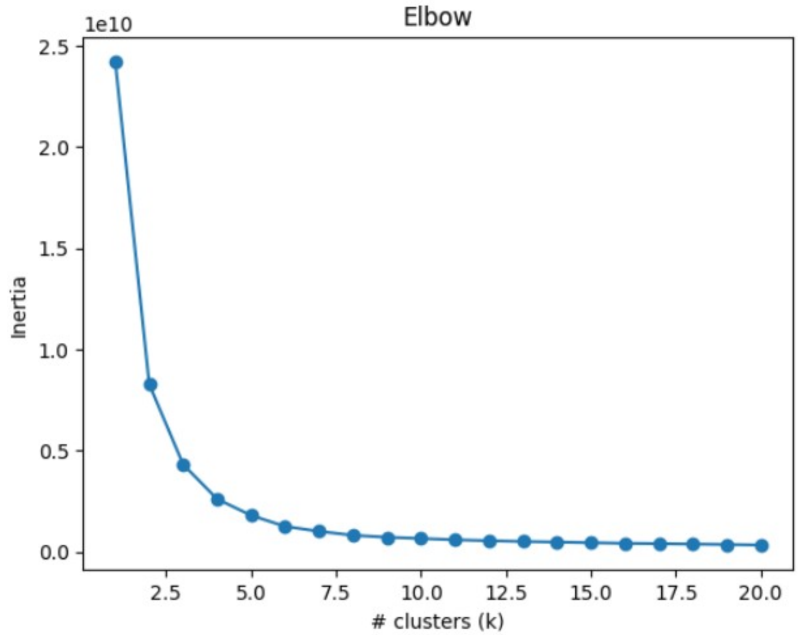
Additional Features

While the data set includes quite a lot of information about the activity of the participants, it lacks many features that could be useful in prediction. It's possible that our predictive ability was weak because we lacked essential features to control for, including demographic data (such as age and gender) and psychographic data (such as lifestyle habits) of the users.

Appendices

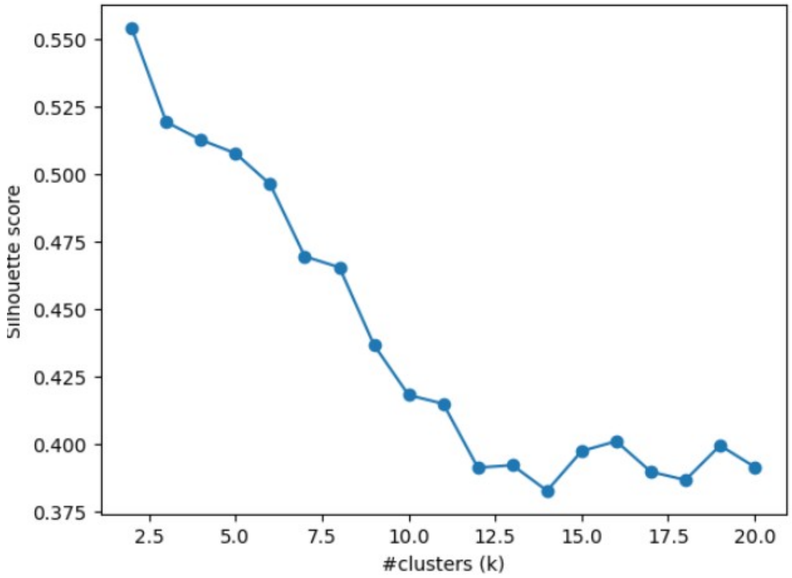
Appendix 1: K-Means Clustering Elbow and Silhouette Curves

Figure A1: Elbow Curve



Elbow method Results:
k=2: 8277287813.8112755
k=3: 4312097945.80378
k=4: 2607087929.4646845
k=5: 1796748742.0001087
k=6: 1254136855.3887343
k=7: 1011676613.0721545
k=8: 815614698.7685683
k=9: 719938009.0535412

Figure A2: Silhouette Curve



Silhouette Score results:
k=2: 0.5541558375276258
k=3: 0.5175001046048336
k=4: 0.5127001327488621
k=5: 0.48882697288531457
k=6: 0.49651030957911
k=7: 0.47578118869473296
k=8: 0.4394096513225059
k=9: 0.43674114408378106

Appendix 2: Detailed Observations of K-Means Clusters

Cluster 1		StepTotal	Calories	SleepMin	ActiveMin
This cluster contains data from over a total of 185 days. On average, each day in this cluster saw a StepTotal of around 6,111 and total Calories burned of approximately 2,203. Sleep duration, or SleepMin, averaged 188 minutes per day, with a standard deviation of nearly 237 minutes indicating quite a range. Meanwhile, time spent in ActiveMinutes averaged around 244 minutes daily. Some notable patterns emerge. While StepTotal and Calories indicate a moderately active lifestyle on average, the lower than average SleepMin of 188 minutes, compared to normal sleep needs, suggests a mistake in the measurement. Interestingly, days with less sleep tended to see higher ActiveMinutes, around 244 on average. This possible correlation hints that on days with less rest, more physical activity was undertaken by these individuals. However, it's important to note that since the data represents days across multiple people, differences in lifestyle and schedules may influence these metrics for certain individuals within the cluster. Overall, this cluster shows intriguing connections between sleep, activity levels, and other daily variables when explored at an aggregate, daily level across a group of people.	count	185	185	185	185
	mean	6110.72973	2202.66487	188.2	243.85405
	std	793.72856	514.61571	236.93057	77.18925
	min	4732	1376	0	0
	25%	5372	1820	0	196
	50%	6116	2012	0	243
	75%	6805	2681	425	295
	max	7475	3645	741	416

Cluster 2		StepTotal	Calories	SleepMin	ActiveMin
This cluster contains data aggregated from 106 days. On average, each day saw a StepTotal of around 14,311 and Calories burned totaling approximately 2,782. Average SleepMinutes was 208, with a standard deviation of nearly 216 minutes showing variability. ActiveMinutes averaged 328. A key aspect of this cluster is the high levels of physical activity, as seen in the StepTotal and ActiveMinutes averages. However, the SleepMinutes mean of 208 minutes, slightly above the average needed, along with instances of zero sleep recorded raise questions. It makes one wonder if sleep tracking was inconsistently logged by some individuals. Closer examination hints at an inverse relationship between ActiveMinutes and SleepMinutes. On days with higher than average physical engagement of around 328 ActiveMinutes, SleepMinutes tended to decrease. This cluster provides a contrasting picture to the first, with individuals expending substantially more daily energy but allotting marginally less time for sleep. Further exploration of the complex interplay between activity levels, caloric needs, and sleep patterns within this group could offer valuable insight. Overall, high activity levels but variable sleep metrics characterize this cluster.	count	106	106	106	106
	mean	14310.5566	2781.59267	208.17925	328.23585
	std	1151.27371	683.92490	215.52593	81.01353
	min	12685	1620	0	0
	25%	13283.5	2178.75	0	285
	50%	14296	2836.5	104	333
	75%	15116.5	3138.5	414.25	375.5
	max	17076	4392	620	540

Cluster 3		StepTotal	Calories	SleepMin	ActiveMin
Represents 162 days with predominantly restful yet nourishing routines. On	count	162	162	162	162

average, each day included around 283 steps, burning roughly 1,816 calories through a balanced combination of movement and rest. Most days involved 59 minutes of sleep, which leads to a question about the accuracy of the measurement. Minimal time was devoted each day to high-intensity exercise according to the average of 24 active minutes.	mean	282.93209	1816.33577	59.27777	23.913580
	std	527.73215	419.29986	152.9907	48.815331
	min	0	65	0	0
	25%	0	1488	0	0
	50%	0	1992	0	0
	75%	244	2098.5	0	33.75
	max	1892	2668	840	297

Cluster 4		StepTotal	Calories	SleepMin	ActiveMin
Represents 166 days of heightened hustle and bustle. On average, each day saw over 11,000 steps burn about 2,540 calories through nearly constant motion. Most allotted 291 minutes for sleep with a median of 293 active minutes.	count	166	166	166	166
	mean	11062.6686	2540.6858	291.66265	287.18072
	std	819.263204	659.33632	218.37983	84.637787
	min	9837	1492	0	0
	25%	10328	2036.75	0	255.25
	50%	10918	2478	391.5	293
	75%	11704	2964.5	473	330.25
	max	12669	4502	640	513

Cluster 5		StepTotal	Calories	SleepMin	ActiveMin
Cluster 5 depicts 29 days. On average, each individual tackled over 19,000 steps while burning around 3,440 calories daily suggesting a high-intensity rate. Accordingly, the activity levels rarely dipped below 333 minutes. Aligned with other clusters, a mere 73 minutes were tracked in the sleeping time with instances of no rest at all. A pattern that strongly suggests damaged trackers or that the subjects did not take the measurements properly. Thus, a considerably high activity rate is evident, surpassing the average of the other clusters and suggesting a heightened pace of activity with significant calorie burning.	count	29	29	29	29
	mean	19649.7642	3439.75862	73.03448	332.65517
	std	2911.11697	455.57700	163.00865	108.07380
	min	7370.1625	2159	0	0
	25%	18229	3211	0	311
	50%	20031	3555	0	334
	75%	21391	3790	0	381
	max	23186	4022	539	483

Cluster 6		StepTotal	Calories	SleepMin	ActiveMin

In this cluster, individuals did around 7,370 steps while expending about 3,107 calories. Consistent with the earlier clusters, the pattern of inadequate sleep tracking persists, averaging a mere zero minutes of sleep, while activity spans consistently between 352 and 552 minutes. In this cluster, the absence of sleep time is evident with 0 in almost every measure of this metric, raising an alert for further analysis	count	3	3	3	3
	mean	7370.1625	3107.27430	0	443.33333
	std	1.11E-12	1131.75022	0	101.12039
	min	7370.1625	2241.82291	0	352
	25%	7370.1625	2466.91145	0	389
	50%	7370.1625	2692	0	426
	75%	7370.1625	3540	0	489
	max	7370.1625	4388	0	552

Cluster 7		StepTotal	Calories	SleepMin	ActiveMin
Within this cluster of 151 days, the mean StepTotal is 3,311 steps. This falls below the average but above the minimum recorded value of 1,758 steps, indicating some light daily physical activity. The mean Caloric intake is 1,924 calories per day, again slightly below average but well above the minimum of 716 calories. Sleep duration averages 231 minutes or just under 4 hours per night, with a standard deviation of 264 minutes suggesting quite variable sleep patterns across individuals. Time spent in actively moving minutes averages 148 minutes per day. This cluster reflects a moderately sedentary lifestyle with light physical activity and variable sleep.	count	151	151	151	151
	mean	3311.43046	1924.68211	231.73509	148.21192
	std	830.65035	465.58819	264.02155	56.15770
	min	1758	716	0	0
	25%	2568.5	1561.5	0	117.5
	50%	3403	1880	51	144
	75%	3959	2311.5	489.5	191.5
	max	4747	3059	928	291

Cluster 8		StepTotal	Calories	SleepMin	ActiveMin
This cluster contains 158 days. The mean StepTotal is 8,649 steps, well above the average and close to the maximum recorded value of 9,834 steps per day. This reflects a highly active lifestyle with extensive daily walking or other physical activity. Caloric intake averages 2,439 calories, again above average and indicative of sufficient nutrition to support the elevated activity levels. Sleep duration averages 215 minutes or just over 3.5 hours per night, with a standard deviation of 229 minutes showing inconsistent sleep patterns. Actively moving minutes are the highest of all clusters at an average of 290 minutes per day, highlighting substantial daily physical exertion. This cluster portrays an extremely active lifestyle with moderate sleep and higher caloric needs to fuel extensive daily activity.	count	158	158	158	158
	mean	8649.86708	2439.80379	215.02531	289.72784
	std	699.30757	597.82088	228.7838	93.99076
	min	7359	1297	0	0
	25%	8067.5	1983.5	0	244
	50%	8596.5	2249	75	290.5
	75%	9228.25	2938.75	443.25	353.75
	max	9834	4097	679	548

Appendix 3: Hierarchical Clustering Dendrogram and Silhouette Curve

Figure A3: Hierarchical Clustering Dendrogram

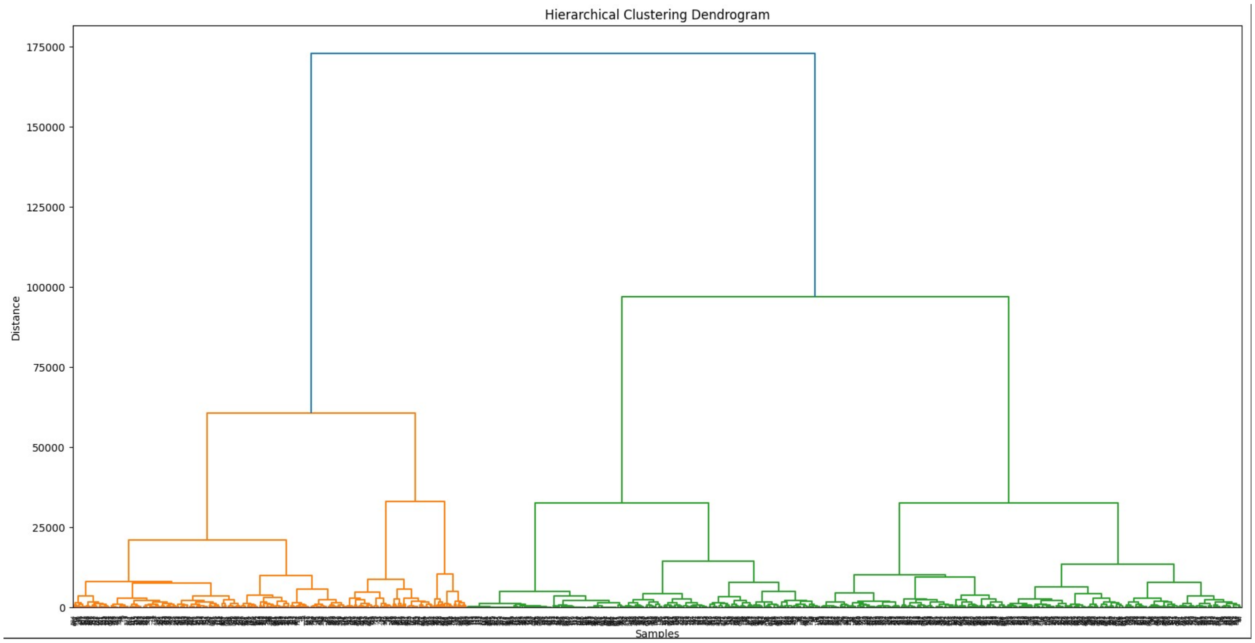
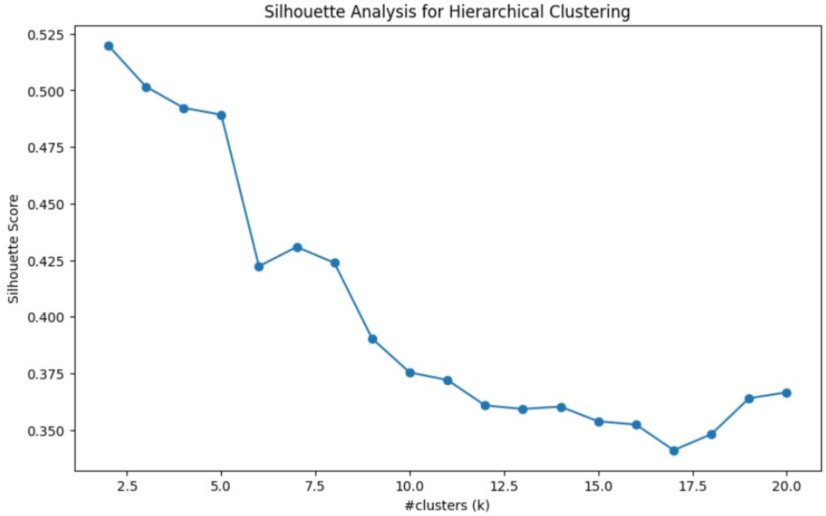


Figure A4: Silhouette Curve



Silhouette Score results:

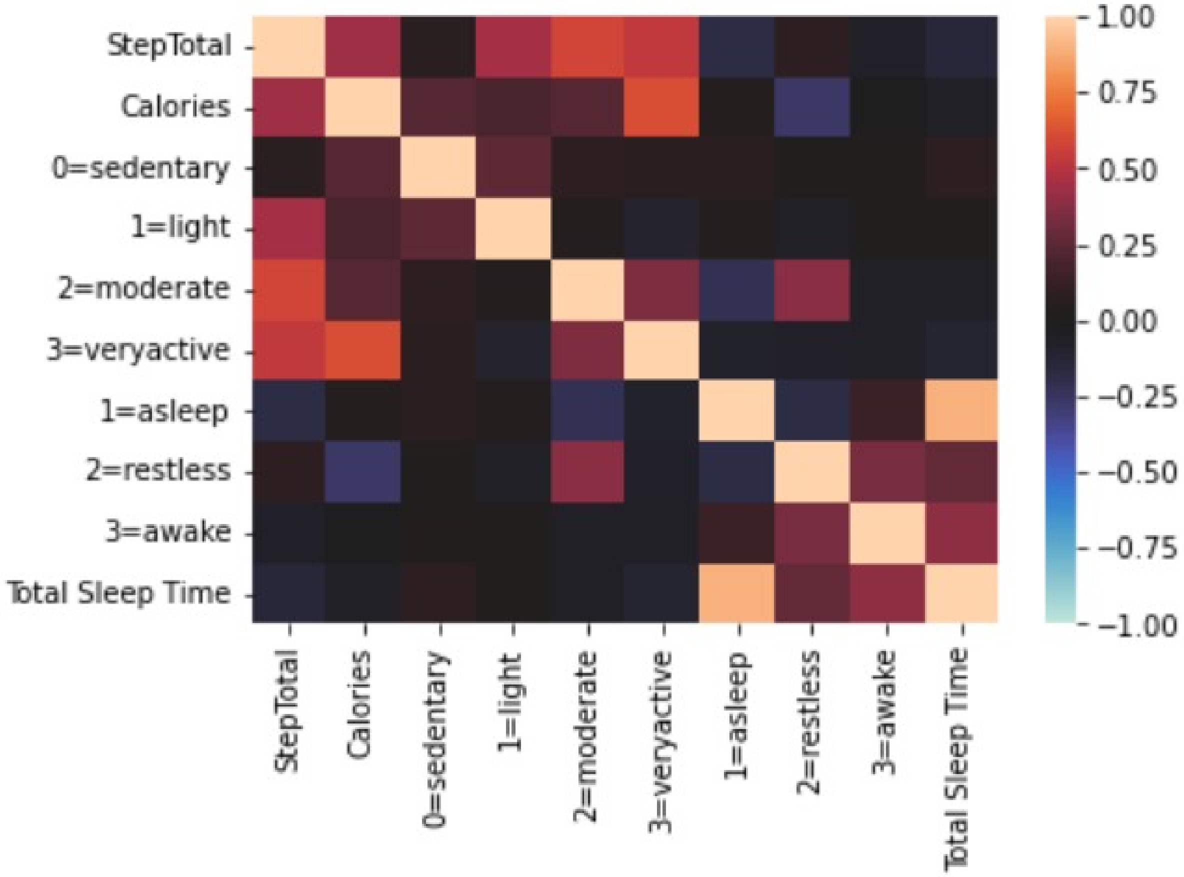
k=2: 0.5198727203604108
k=3: 0.5016244032946513
k=4: 0.4922521284396789
k=5: 0.4892154211751792
k=6: 0.42226516479645254
k=7: 0.4308087842417816
k=8: 0.42379781976063785
k=9: 0.3903798167630606

Appendix 4: Detailed Description of Hierarchical Clustering Clusters

Cluster 1		StepTotal	Calories	SleepMin	ActiveMin
In the first cluster, comprising 323 observations, various metrics are examined to understand their impact on sleeping activity. The mean StepTotal for this cluster is approximately 12,803, with a standard deviation of 2,951. This suggests that observations in this cluster tend to be moderately active throughout the day. The mean Caloric intake is around 2,706, implying a relatively balanced energy intake. The mean SleepMin, which represents the duration of sleep, is approximately 242 minutes (around 4 hours), with a notable standard deviation of 222 minutes. This suggests a considerable variability in sleeping patterns within this cluster. When considering ActiveMin, with a mean of about 303 minutes, it appears that observations in this cluster engage in physical activity for a substantial amount of time. Notably, the minimum value of SleepMin is 0, indicating instances where observations might not have recorded any sleep. The median (50th percentile) SleepMin is 324 minutes, showing that half of the observations sleep for more than 5 hours. Interestingly, the maximum SleepMin recorded is 679 minutes, suggesting that some observations sleep for nearly 11.5 hours.	count	323	323	323	323
	mean	12803.57	2706.47	242.31	302.954
	std	2951.2715	684.195	222.23	86.5372
	min	9143	1492	0	0
	25%	10538	2107	0	264.5
	50%	12159	2757	324	303
	75%	14296	3151.5	449.5	345.5
	max	23186	4392	679	540

Cluster 2		StepTotal	Calories	SleepMin	ActiveMin
The second cluster includes 637 observations, for whom sleeping activity metrics are analyzed. The mean StepTotal for this cluster is notably lower at around 4,478, with a standard deviation of 3,091. This indicates a less active lifestyle compared to Cluster 1. The mean Caloric intake is approximately 2,089, suggesting a relatively moderate energy intake. The mean SleepMin for this cluster is about 169 minutes (approximately 2.8 hours), with a higher standard deviation of 233 minutes. This variability might indicate inconsistent sleeping patterns among observations in this cluster. Interestingly, a considerable portion of the cluster has recorded 0 minutes of sleep, which could indicate missing data or instances of observations not tracking their sleep. The mean ActiveMin is around 177 minutes, indicating a lower engagement in physical activity compared to Cluster 1.	count	637	637	637	637
	mean	4478.41	2086.92	169.11	176.97
	std	3091.93	546.35	233.19	126.23
	min	0	247	0	0
	25%	1727	1725	0	62
	50%	4880	2016	0	187
	75%	7162	2362	409	266
	max	9799	4388	928	552

Appendix 5: Correlation Matrix Heat Map for Classification Model Data


StepTotal
Calories
0=sedentary
1=light
2=moderate
3=veryactive
1=asleep
2=restless
3=awake
Total Sleep Time