FitBit Data Analysis

Machine Learning
Group 8

Julia Ciriello
Nik Gupta
Devin Ridley
Alvaro Rivera-Eraso

Executive Summary

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. 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.

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.

Clustering Models

For the clustering stage, two models have been considered: a K-Means model and a hierarchical clustering model. The features under consideration are StepTotal, Calories, SleepMin, and Active Min. To determine the optimal number of clusters for the K-Means model we used the elbow method, silhouette analysis, and GridSearchCV. These hyperparameter tuning methods yielded different results so we proceeded with the output of the grid search, k=8 clusters (Appendix 1). 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.

The hierarchical model started with a similar approach which yielded the optimal number of clusters as k=2 (Appendix 2). 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.

Classification Models

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. The target boolean variable "Good Sleep" was defined as whether a person slept for at least 7 hours in a given night. 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. 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 (Appendix 3).

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 (Appendix 4) 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

Both clustering models performed reasonably well. The K-means model demonstrated slightly inferior performance to hierarchical clustering. The hierarchical clustering model took a more suitable approach in generating group distribution. 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. 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.

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.

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. These include: an increased number of participants and longer study duration; complete and consistent weight data; additional demographic and psychographic user data. These factors have the potential to significantly improve our models' performance.

Appendices

Appendix 1: 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.

Appendix 2: 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.

Appendix 3: 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**

Appendix 4: Correlation Matrix Heat Map for Classification Model Data