Analyzing Daily Activity and Sleep Patterns

CS-C4100 - Digital Health and Human Behavior
Course Project

Contents

1 Introduction

Wearable devices like Fitbit are widely used in tracking daily physical activities, sleep patterns, and overall health. Bisson et al. (2019) utilized Fitbit Zip devices to track steps and active minutes and then assessed self-reported sleep quality and duration. They found in their study that Individuals, especially women, with higher average active minutes reported better sleep quality across the month. This connection is illustrated in figure 1.

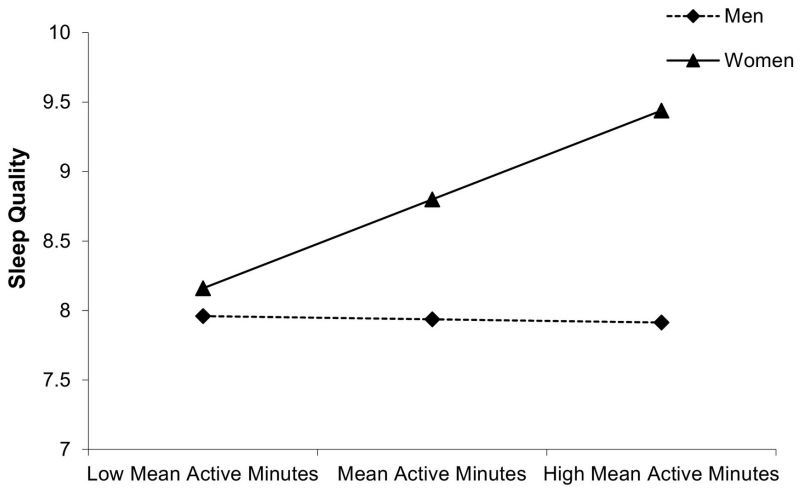

Figure 1: Relationship of mean daily active minutes and sleep quality across the month by sex.
(image from Bisson et al. (2019))

Another related study was conducted by Lakerveld et al. (2016). It highlights an association between short sleep duration and increased screen time, but it could not find a link between sleep duration and other sedentary behaviours. The study suggests that reducing screen time could lower sedentary behaviour and improve sleep duration, potentially reducing obesity and related health risks.

There are also several other studies implicating the health benefits of physical activity, and that have used data gathered with fitness tracker devices. According to Bassuk & Manson (2005), physically active individuals have a 30–50% lower risk of developing type 2 diabetes compared to sedentary individuals. They also noticed that moderate physical activity, such as brisk walking for 30 minutes a day, provides metabolic and cardiovascular benefits.

This study aims to find trends and relationships between activity and sleep metrics, using Fitbit tracker data. This process includes making observations at

both individual and group levels, clustering the data to categorize living styles and use of regression models to predict the amount of time spent asleep after certain time of activity. The data in this study is acquired from Kaggle (*FitBit Fitness Tracker Data — kaggle.com* 2024). Information like this is valuable when considering how daily habits affect health outcomes. It can also be used when designing fitness trackers and tracker device features, such as predicting sleep quality outcomes based on physical activity.

Findings achieved in the analysis show correlations between active minutes and sleep duration, variations in activity and sleep by weekday, and distinct lifestyle clusters representing healthy and unhealthy behaviours. This report is divided into four main sections. The first section explains the current problem under analysis. The second section describes the dataset used in this study, and the third section presents the methods used in the data analysis. The last section discusses the results and concludes the study.

2 Problem Formulation

The primary focus of the analysis is to investigate the relationship between daily activity and sleep patterns as recorded by Fitbit devices. The ultimate goal is to understand how physical activity, sedentary behaviour, and energy consumption can affect the duration of sleep. Even though there are many studies related to wearable fitness trackers, this specific topic between physical activity, sedentary behaviour, and sleep duration is not widely studied.

In analysing activity and sleep patterns, a few approaches will be used. The first approach is to compare trends within individuals and across the data as a whole to generalize findings and discover behavioural patterns. After this, a clustering analysis is conducted. At this point, k means clustering is used to group participants into separate lifestyle clusters, in hopes of identifying distinct patterns of healthy and unhealthy habits. Lastly, two different regression models, random forest and gradient boosting, are built to predict the time spent asleep.

The primary objectives of this study are summarised below:

- Clean and preprocess the data, addressing possible missing values and outliers.
- Analyse life habits at both the individual and group levels.
- Compare daily activity metrics such as activity minutes, daily calories, step counts, and sleep durations between each other and possibly with some recommendation reference values (such as WHO's suggested reference values).
- Find if there is a dependency between the activity and sleep duration.
- Model the dependency between the activity and sleep duration with clustering and regression models.

3 Dataset

The dataset used for this project is acquired from the Fitbit data repository on Kaggle (*FitBit Fitness Tracker Data — kaggle.com* 2024). It includes multiple CSV files from different tracking domains and with different resolutions (data captured daily, hourly, and by minute). The data is captured over a month. To prevent the study from scaling too large only two datasets were picked from the repository. The two datasets and the merged version of these two are briefly described in the following parts. All the data except dates (used as index) are numerical.

3.1 DailyActivity_merged

DailyActivity_merged contains the information of the daily activities of the user. This includes features such as the number of steps taken, distance moved, active and sedentary minutes, and calories burned. The data does not contain any missing values. In the dataset, there are 940 entries and 33 different attendants. This dataset's main features are further described in table 1.

Table 1: Main features of the dataset dailyActivity_merged.

Feature Name	**Mean**	**Std**	**Min**	**Max**
TotalSteps	7637.9	5087.2	0	36019
TotalDistance	5.5	3.9	0.0	28.0
TrackerDistance	5.5	3.9	0.0	28.0
LoggedActivitiesDistance	0.1	0.6	0.0	4.9
VeryActiveDistance	1.5	2.7	0.0	22.0
ModeratelyActiveDistance	0.6	0.9	0.0	6.5
LightActiveDistance	3.3	2.0	0.0	10.7
SedentaryActiveDistance	0.0	0.0	0.0	0.1
VeryActiveMinutes	21.2	32.8	0.0	210.0
FairlyActiveMinutes	13.6	20.0	0.0	143.0
LightlyActiveMinutes	192.8	109.2	0.0	518.0
SedentaryMinutes	991.2	301.3	0.0	1440.0
Calories	2303.6	718.2	0.0	4900.0

3.2 SleepDay_merged

SleepDay_merged contains information on total minutes asleep per night, total time spent in bed, and number of separate sleeping records per night. It is gathered from 24 individuals. The main features of the dataset are described in table 2. The data contains 413 entries. The data does not contain any missing values.

Table 2: Main features of the dataset sleepDay_merged.

Feature Name	**Mean**	**Std**	**Min**	**Max**
TotalMinutesAsleep	419.5	118.3	58.0	796.0
TotalTimeInBed	558.6	127.1	61.0	961.0
TotalSleepRecords	1.1	0.3	1	3

3.3 Final Merged Dataset

The final dataset used in the further analysis and in modelling the dependency between activity, calories and total time asleep was acquired by merging the two aforementioned datasets by participant id and date. The merged dataset consisted of 397 rows and 17 columns. After merging the two datasets few new features were engineered from the existing ones. These new features were "SumActiveMinutes", "ModerateActiveMinutes", and "TotalMinutesAwakeInBed". SumActiveMinutes was acquired by taking the sum of active minutes of a row, that is, the sum of all activity minutes except sedentary minutes. ModerateActiveMinutes was created by taking the sum of the VeryActiveMinutes and FairlyActiveMinutes. TotalMinutesAwakeInBed is the difference between the TotalTimeInBedTotalTimeInBed and TotalMinutesAsleep. Figures 2 and 3 show the distribution of activity and sleep-related features in the final merged dataset. Further breakdown of the data preprocessing is discussed in the following part.

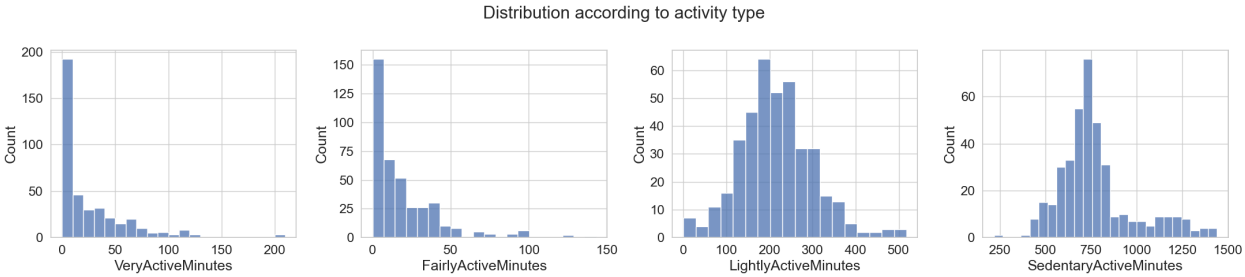

Figure 2: Activity type distributions in the final merged dataset.

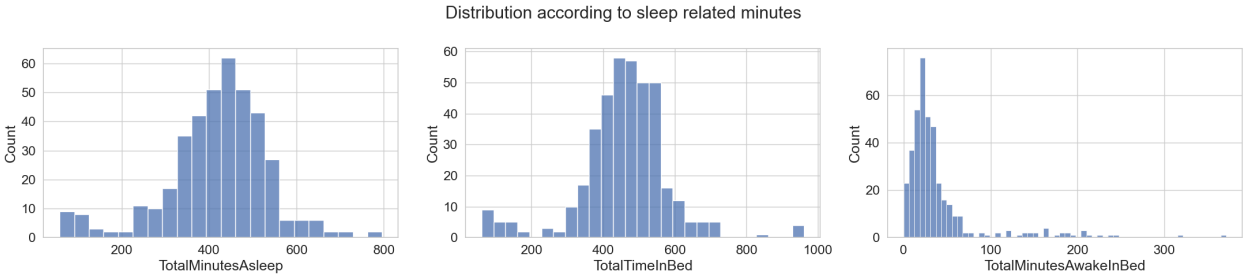

Figure 3: Sleep-related features' distributions in the final merged dataset.

4 Methods

This section describes the methods used to analyse the Fitbit data. The methodology consisted of three primary phases: data preprocessing and exploratory analysis, clustering analysis, and regression modelling.

4.1 Data Preprocessing and Exploratory Analysis

The first step in the analysis involved data preprocessing. Datasets consisting of daily activity and sleep records were loaded and merged. At first, it was challenging to align sleep records with daily activity data, as sleep records often span two calendar days. To solve this, the dates in the sleep dataset were shifted by one day to correspond to the preceding day's activity. In this phase tools such as Scikit-learn (Pedregosa et al. 2011), Numpy (Harris et al. 2020), Pandas (pandas development team 2020), and Seaborn (Waskom 2021) were used.

To enhance the analysis, additional features were engineered. The SumActiveMinutes variable was calculated by summing all activity minutes except the sedentary time. ModerateActiveMinutes was created by taking the sum of the VeryActiveMinutes and FairlyActiveMinutes. TotalMinutesAwakeInBed was derived by subtracting total sleep time from the time spent in bed. Data cleaning was also conducted to remove outliers. A couple of possible outlier values were found in TotalMinutesAwakeInBed, which could distort results so they were removed. Finally, to prepare the data for statistical and machine learning methods, numerical variables were standardized using Scikit-Learn StandardScaler, that is, the z-score for each feature was computed.

To capture overall lifestyle patterns, participant-level data was also aggregated by averaging daily metric features such as SumActiveMinutes, SedentaryMinutes, TotalMinutesAsleep, and Calories. The aggregation resulted in a dataframe that contained only one row entry for each subject so that one row contained averaged metrics of a subject from the whole month period. This aggregated version of the data was used for clustering only.

To better understand the correlations between all the features, a correlation matrix was plotted. This correlation matrix can be seen in figure 4. The matrix shows that activity-engaged minutes have only a slight positive correlation with the total minutes asleep feature.

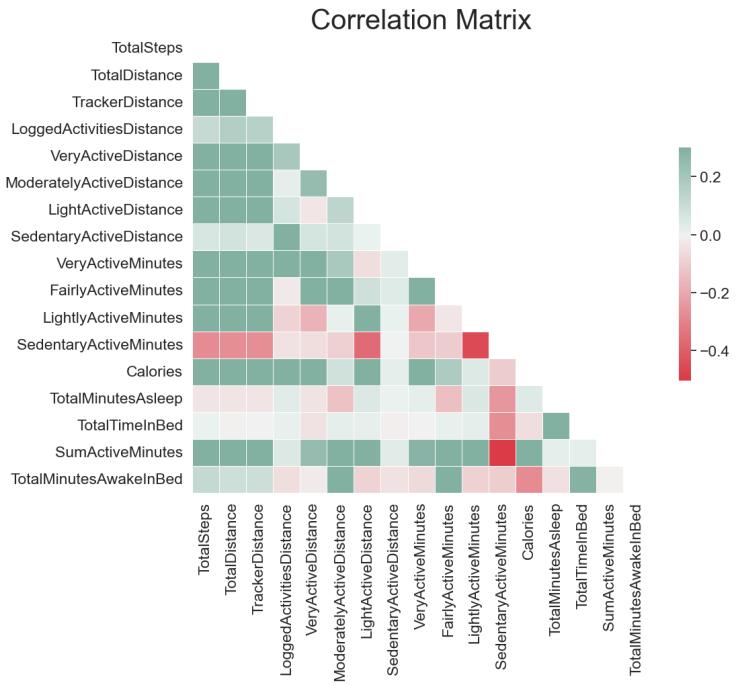

Figure 4: Correlation matrix for the features at the group level.

4.2 Clustering

Clustering was conducted to identify different lifestyle patterns among participants. Features used for clustering included SumActiveMinutes, TotalMinutesAsleep, Calories, and SedentaryMinutes. To determine the optimal number of clusters, K-means clustering was tried with varying values of k (from 2 to 8). At every iteration of k, the clustering results were evaluated using the silhouette score and Calinski-Harabasz index. The clustering was done both to cluster row entries of the data (using the scaled data) and also to cluster the individuals (using the individual-level aggregated scaled data). The silhouette score and Calinski-Harabasz indices were plotted from each iteration, and the optimal value of k was decided by looking at the first shared "elbow" point in the two plots.

Based on these metrics, a value of $k = 3$ was chosen for the day-level cluster analysis and $k = 4$ *and* $k = 2$ were chosen for the individual-level cluster analysis. Visualizations, including boxplots and pie charts, were used to illustrate the clusters' feature values.

4.3 Regression models

To model the relationship between activity, calories, and sleep metrics two regression models were developed. These models predicted total minutes spent asleep based on moderate activity minutes, sedentary time, and calories burned from the day before the night. To compare the performance of the two regression models a robust evaluation setup was implemented using 5-fold cross-validation. In this approach, the dataset was split into five subsets. Then, iterating all five different folds, the model was trained on four folds. At each iteration, the models generated out-of-fold predictions on the fifth fold, ensuring that every data point was used for both training and testing. The difference in the accuracy between the two models was tested using the Wilcoxon signed-rank test. Eventually, the random forest model was chosen for further analysis. In further analysis, the dataset was split into training and testing subsets, and performance was evaluated using metrics such as Mean Squared Error (MSE) and R^2. Also impurity-based feature importance and predicted values were plotted when analysing the random forest model. The results are further discussed in the next part.

5 Results

The results suggest that the pattern between the activity and sleep duration does exist, although this should be more adequately tested. This means considering the role of the other features in the data, and the use of several hypothesis testing methods. The following sections further discuss the results obtained in different phases of the analysis. The limitations are discussed in the final section.

5.1 Visualisations

Visualisations were made both on the individual subject-level data and on the group-level data. In both of these phases, non-scaled and scaled data were used. The individual-level analysis was conducted on the participant with the most complete records, with a total of 30 records.

5.1.1 Observations on the Subject Level

After preprocessing and scaling, the analysis showed variations in activity and sleep behaviours across the different days of the week. The individual subject under analysis displayed reduced activity levels on weekends compared to weekdays but tended to sleep for longer durations on these days. Boxplots were generated to visualize the distributions of active minutes and sleep durations by weekday. Figure 5 shows how the participant's weekly schedule was structured, with higher levels of activity and shorter sleep durations on weekdays and recovery days on the weekends.

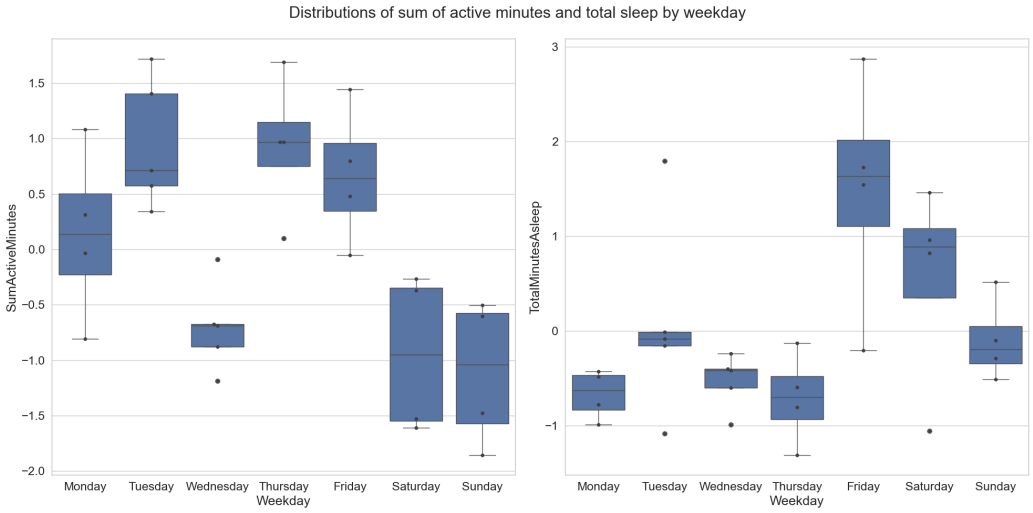

Figure 5: Individual subject's activity and sleep distribution by weekday (using standard-scaled features).

5.1.2 Observations on the Group Level

At the group level, the analysis showed broader trends. Figure 6 contains a histogram of daily sleeping records (ignoring subject id information) and it colour codes how some nights the sleeping recommendation (Watson et al. 2015) is met and sometimes not. One can see how most of the nights individuals meet the recommendation of at least seven hours of sleep per night.

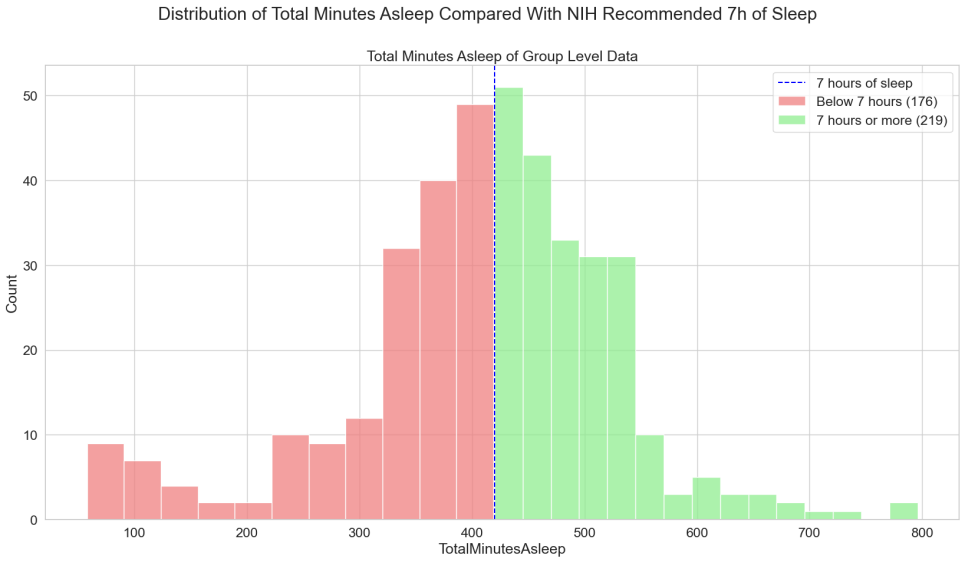

Figure 6: Group-level sleep duration distribution.

Interestingly, as can be seen in figure 3, many participants spent considerable time awake in bed, indicating poor "sleeping efficiency" for many individuals.

5.2 Clustering

Clustering analysis was conducted in two distinct phases: clustering of individual days (using $k = 3$) and clustering of aggregated participant data (using $k = 4$ and $k = 2$).

5.2.1 Clustering day entries

In the first phase, K-means clustering was applied to daily records, focusing on features such as SumActiveMinutes, SedentaryMinutes, TotalMinutesAsleep, and Calories. Based on silhouette scores and Calinski-Harabasz indices, $k = 3$ was selected as the optimal number of clusters. These clusters captured three distinct types of daily behaviors.

First cluster clearly highlighted active days represented by high levels of physical activity and relatively low sedentary time. These days were also associated with moderate sleep durations and higher calorie expenditure.

The second cluster could be seen highlighting somewhat balanced days, characterized by a moderate amount of activity and sedentary behavior. These days had sufficient sleep durations and calorie expenditure, reflecting a well-rounded lifestyle.

The third cluster consisted of a lot of sedentary-oriented days. These days contained long sedentary periods, minimal physical activity, insufficient sleep, and the lowest calorie expenditure among the clusters.

Visualizations of the feature distributions across clusters revealed clear separations. As can be seen in figures 7 and 8, the "Active Days" cluster had the highest SumActiveMinutes, while the "Sedentary Days" cluster exhibited significantly higher SedentaryMinutes.

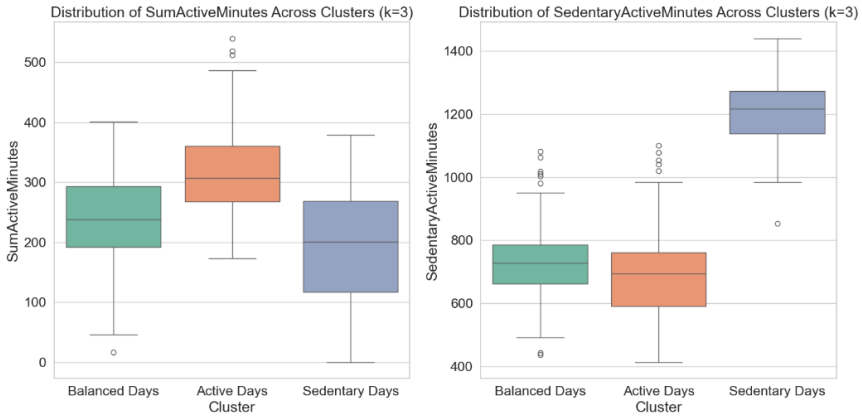

Figure 7: Active and sedentary minutes a day among the three cluster types.

The characteristics in these three clusters seem to make ground to the idea that days with more activity are followed by nights with longer sleep duration. This clustering was done to the whole data disregarding individual subject ids. The next part discusses how clustering was done with subject individuals, possibly showing how individual's behaviours can be observed and categorised.

5.2.2 Clustering individual subjects

To capture overall lifestyle patterns, participant-level data was aggregated by averaging daily metrics. Features such as SumActiveMinutes, SedentaryMinutes, TotalMinutesAsleep, and Calories were used for clustering. After scaling the data, two optimal number of clusters were considered. It seemed that $k = 4$ clusters would

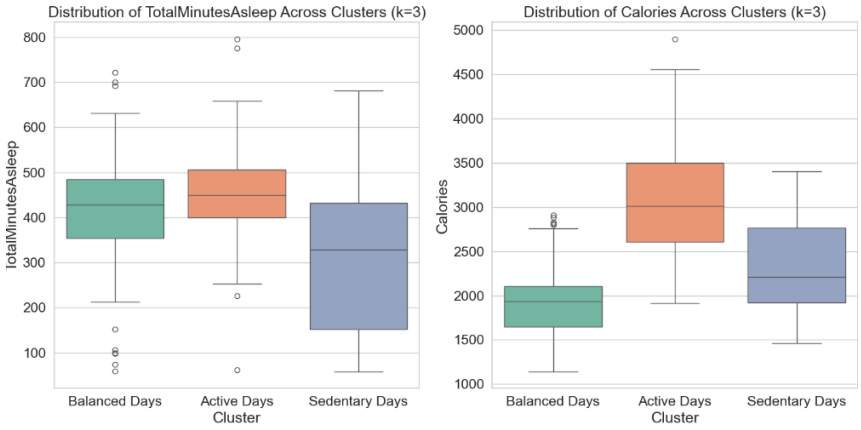

Figure 8: Minutes asleep a night and calories burnt per day among the three cluster types.

have been the most optimal based on the same clustering metrics as used when clustering day-level entries. Second optimal solution was to choose $k = 2$. To explore simpler patterns, the aggregated data was clustered into $k = 2$ groups. This approach divided participants into two broad categories:

Active and Balanced Individuals: Participants in this group exhibited higher activity levels, better sleep durations, and lower sedentary time. Their calorie expenditure was significantly higher.

Sedentary Individuals: This group was characterized by lower activity levels, extended sedentary periods, and inadequate sleep durations. Calorie expenditure was also comparatively low.

The study then further analysed how individuals in the two clusters meet the standard recommendation of 7h of sleep per night for adults and also a 42 minutes of moderate-intensity physical activity a day (derived from a recommendation of 300 mins a week) (WHO 2020). To achieve this the data was first re-aggregated to obtain also the average of the sum of the veryActiveMinutes and FairlyActiveMinutes per individual. The sum of the veryActiveMinutes and FairlyActiveMinutes per individual represents the moderately active time spent by an individual. Then pie charts were plotted depicting how well the recommendations are met within the two clusters. Figures 9 and 10 show, how majority of the individuals in cluster 0 meet the recommended sleeping time while majority in cluster 1 do not.

Similar behaviour among the two clusters can also been seen in physical activity. Figures 11 and 12 show how very small portion of individuals in cluster 1 meet the 42min physical activity a day threshold while almost half of the individuals do meet this in cluster 0.

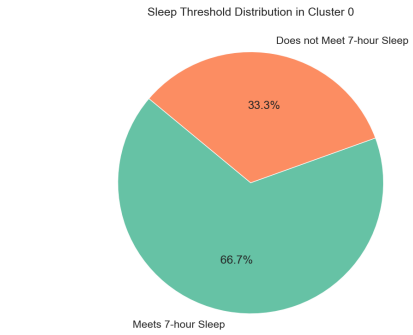

Figure 9: Individuals in cluster 0 that meet or do not meet the recommended 7h sleep.

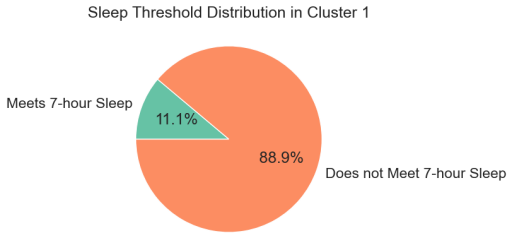

Figure 10: Individuals in cluster 1 that meet or do not meet the recommended 7h sleep.

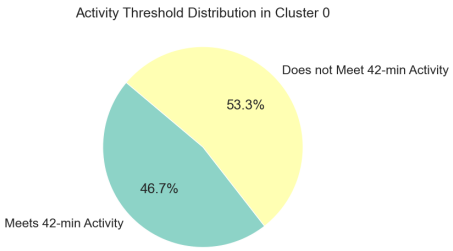

Figure 11: Individuals in cluster 0 that meet or do not meet the recommended 42 mins. moderate physical activity.

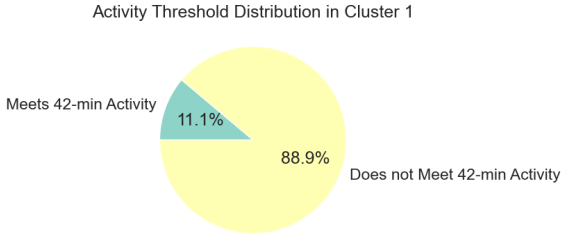

Figure 12: Individuals in cluster 1 that meet or do not meet the recommended 42 mins. moderate physical activity.

5.3 Regression models

The random forest regressor and gradient boosting regressor both had moderate predictive performance. However, the gradient boosting regressor produced lower RMSE values across the folds, indicating slightly better accuracy in predicting total minutes spent asleep. The difference in the accuracy between the two models was tested using Wilcoxon signed-rank test, which indicated that the difference in performance between the two models was not statistically significant (with 95 % confidence level).

Eventually the random forest model was choosed to further analysis, since the random forest model had a slightly better RMSE and it is easier to interpret (e.g., feature importance visualization). Figure 13 shows a scatter plot of the predicted values together with the actual target values. The red line in the figure demonstrates the so called "perfect fit", indicating that the closer the points are to the red line, the better the fit of the model.

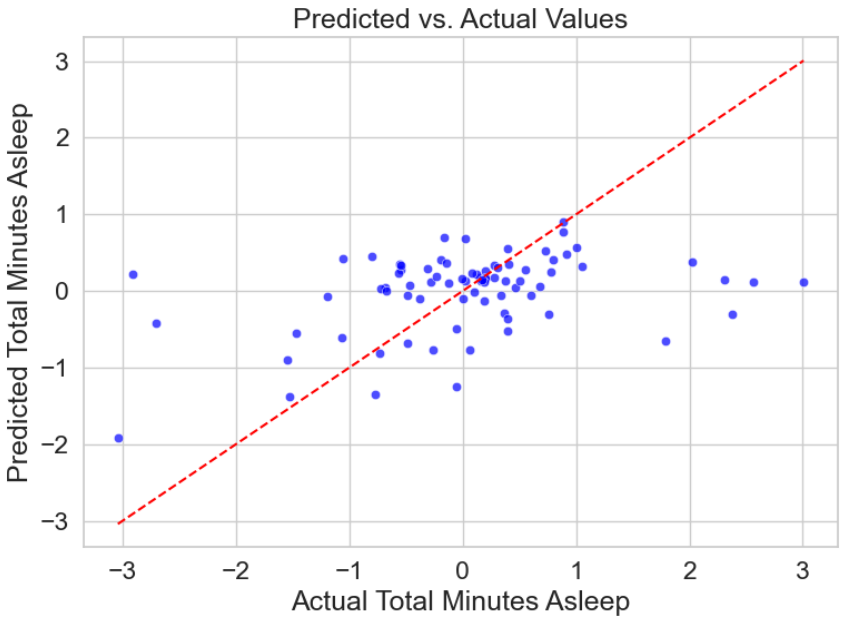

Figure 13: Scatter plot of the actual values and predictions made by the random forest model (using scaled values of the target "TotalMinutesAsleep").

In figure 14 the impurity-based feature importance are plotted for the random forest model. In this figure, one can see that the activity minutes actually had the least importance in the model, and the sedentary minutes had the most. One has to be careful when interpreting these importance results, since the features are most likely linearly related. More on this is discussed in the limitations section.

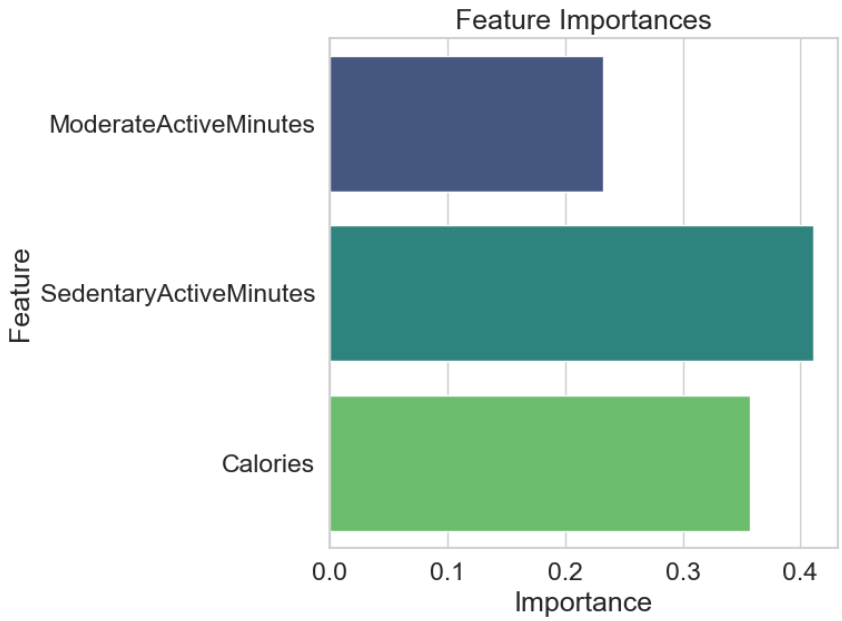

Figure 14: Feature importance extracted from the random forest model.

6 Conclusion and Discussion

The purpose of this study was to analyse healthy and unhealthy behaviours at both group level and at individual subject level. Main focus in these behaviours was in the relationship between activity and sleep time following the activity.

6.1 Limitations and Future Prospects

One has to keep in mind that the Fitbit trackers along with other low-cost activity trackers are prone to lot of measuring errors. Thus one should consider either higher-quality tracking devices or engineer more error-proof designs of data processing when dealing with trackers such as Fitbit.

In a study conducted by Degroote et al. (2020) six low-cost activity trackers were evaluated for the validity of measuring steps, moderate-to-vigorous physical activity (MVPA), and total sleep time (TST) in free-living conditions compared to research-grade accelerometers and armbands. The study showed that low-cost trackers are most accurate in measuring steps, and all tested trackers, including Fitbit, showed weak validity in measuring MVPA. Even though step counts are the most reliable metric across wearable devices, the analysis prioritised the activity minutes over steps when assessing activity patterns. This choice was made since step counts have been already so widely used in several other studies and in the course assignments. The results of the study suggest careful preprocessing and possibly some sort of validation of both activity and sleep data.

In supervised machine learning, one should assess the multicollinearity between the features used. This is because multicollinearity makes it hard to understand how the various features contribute to predicting the target variable (Paul 2006). Considering this, our random forest regression's feature importance must be carefully assessed since the three features used most likely have linear relations. In future, one should assess the multicollinearity and possibly invent mitigations if further developing the regression models,

6.2 Conclusion

This study considered the analysis conducted on the Fitbit tracker data gathered from several persons over a month period. The objectives of this analysis were to preprocess the data appropriately, to analyse life habits at both the individual and group levels, and to model the dependency between the activity and sleep duration with clustering

and regression models. Indeed, several visualisations both on individual and group levels showed patterns between the activity and sleep duration following the activity. The k means clustering managed to separate data into clusters containing data entries with longer activity minutes and longer sleep durations, and vice versa for the other cluster(s). Two regression models, random forest and gradient boosting, were trained to predict sleep duration based on activity time, sedentary time, and calory consumption. The two models showed no statistically significant difference in performance between each other. To conclude, there seems to be pattern between activity levels and following sleep duration but more fine tuned solutions, such as mitigating the multicollinearity between the predictor variables, are needed to further analyse this phenomenon.

References

Bassuk, S. S. & Manson, J. E. (2005), 'Epidemiological evidence for the role of physical activity in reducing risk of type 2 diabetes and cardiovascular disease', *Journal of Applied Physiology* **99**(3), 1193–1204. PMID: 16103522.
URL: *https://doi.org/10.1152/japplphysiol.00160.2005*

Bisson, A. N. S., Robinson, S. A. & Lachman, M. E. (2019), 'Walk to a better night of sleep: testing the relationship between physical activity and sleep', *Sleep Health* **5**(5), 487–494.
URL: *https://www.sciencedirect.com/science/article/pii/S2352721819301056*

Degroote, L., Hamerlinck, G., Poels, K., Maher, C., Crombez, G., De Bourdeaudhuij, I., Vandendriessche, A., Curtis, R. G. & DeSmet, A. (2020), 'Low-cost consumer-based trackers to measure physical activity and sleep duration among adults in free-living conditions: Validation study', *JMIR Mhealth Uhealth* **8**(5), e16674.
URL: *https://doi.org/10.2196/16674*

FitBit Fitness Tracker Data — kaggle.com (2024), https://www.kaggle.com/datasets/arashnic/fitbit. [Accessed 15-11-2024].

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C. & Oliphant, T. E. (2020), 'Array programming with NumPy', *Nature* **585**(7825), 357–362.
URL: *https://doi.org/10.1038/s41586-020-2649-2*

Lakerveld, J., Mackenbach, J. D., Horvath, E., Rutters, F., Compernolle, S., Bárdos, H., De Bourdeaudhuij, I., Charreire, H., Rutter, H., Oppert, J.-M., McKee, M. & Brug, J. (2016), 'The relation between sleep duration and sedentary behaviours in european adults', *Obesity Reviews* **17**(S1), 62–67.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/obr.12381*

pandas development team, T. (2020), 'pandas-dev/pandas: Pandas'.
URL: *https://doi.org/10.5281/zenodo.3509134*

Paul, R. K. (2006), 'Multicollinearity: Causes, effects and remedies', *IASRI, New Delhi* **1**(1), 58–65.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011), 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research* **12**, 2825–2830.

Waskom, M. L. (2021), 'seaborn: statistical data visualization', *Journal of Open Source Software* **6**(60), 3021.
URL: *https://doi.org/10.21105/joss.03021*

Watson, N. F., Badr, M. S., Belenky, G., Bliwise, D. L., Buxton, O. M., Buysse, D., Dinges, D. F., Gangwisch, J., Grandner, M. A., Kushida, C., Malhotra, R. K., Martin, J. L., Patel, S. R., Quan, S. F. & Tasali, E. (2015), 'Recommended amount of sleep for a healthy adult: A joint consensus statement of the american academy of sleep medicine and sleep research society', *Journal of Clinical Sleep Medicine* **11**(06), 591–592.
URL: *https://jcsm.aasm.org/doi/abs/10.5664/jcsm.4758*

WHO, W. H. O. (2020), *WHO guidelines on physical activity and sedentary behaviour*, World Health Organization.