Analysis of calories burn using machine learning techniques

Sai Kiran Ande

*Computer Science University of Central Missouri*

700741579

Taranmeeth Kaur Sardar

*Computer Science University of Central Missouri*

700723388

Yamini Macherla

*Computer Science University of Central Missouri*

700733827

Sri Snigdha Madisetty

*Computer Science University of Central Missouri*

700742420

Git Hub Link: https://github.com/saikiranande3428/Final-Project

*Abstract*—The use of wearable devices that monitor a user's activity, location, and health has experienced a meteoric rise in popularity. It should come as no surprise that adopting a sedentary lifestyle can lead to increased weight gain, the development of chronic and acute illness, and perhaps lower productivity in the classroom, the job, and everyday life in general. Businesses and academic organizations have made significant investments in fitness tracking, with a focus on a wide range of metrics including, but not limited to, sleep quality, cardiovascular health, continuous activity monitoring, general health, illness recovery, and a variety of other measures. Researchers have used deep learning strategies, machine learning algorithms, and a wide variety of statistical methodologies in order to anticipate and recognize the trends, patterns, and deviations in the healthcare data acquired by wearable devices. This has allowed them to predict and understand the data. By examining the data collected from Bella Beat and Fitbit devices, the researchers are hoping to get new knowledge. The data collection for the startup known as Bella Beat is handled by an automated system. In addition, the research investigates the healthcare data gathered by Fitbit in order to identify and investigate any trends, tendencies, or connections that may exist within that data.

There are a lot of privacy problems related with the usage of smart gadgets like the Apple Watch and other similar products, even though these products offer better functionality. The sale of client data is the primary source of revenue for the great majority of well-known technological companies. Each of them is presented with advertisements on their own phones that are pertinent to them as a result of the data acquired by their wearables. Certain people are always concerned with ensuring the safety of their personal space. They stand to gain a great deal from both our research and our model.

***Keywords— calories, prediction, regression, Fitbit, health***

1. INTRODUCTION (*HEADING 1*)

The majority of specialists agree that individuals' day-to- day decisions can have significant effects on their health (such as those related to diet, activity level, and rest). Wearable devices like the Fitbit and the Apple Watch collect vast amounts of data from their sensors in real time. Such information can provide light on a variety of previously unknown health-related habits, such as how often one exercises or how well one sleeps, for example. Since self-

measurement and e-health records do not provide any correlation between the patient's state of health and the wearable's data, the promise of wearables for use in medical research has not been able to reach its full potential. [1]. The amount of heat that is generated by the body as a result of physical activity varies according to a number of different parameters, such as the individual's age, weight, degree of fitness, and the amount of time spent exercising. The goal of this experiment, which will not make use of a calorie counter of any kind, is to determine whether or not certain features of the participants may be utilized to predict the total amount of calories that they burn as a result of the activities that they participate in. Because of this, we have access to a wide variety of equipment, all of which, however, must be worn on the body in order to obtain correct readings from the many internal systems. It is possible to receive real-time updates from an activity tracker if you wirelessly couple it with a mobile device such as a smartphone or tablet. This makes it feasible to receive updates. Because they are so widely available, activity trackers make it possible to keep tabs on one's health as well as their day-to-day activities with the bare minimum of fuss and the utmost comfort. When utilized in the context of health projects, this strategy produces a feedback loop that has the potential to greatly impact efforts to modify behavior. Although it is theoretically conceivable to deduce from these data the current health status of a patient, this avenue of inquiry has not been pursued; rather, the study has concentrated on identifying straightforward relationships[2].

1. MOTIVATION

The vast majority of experts are of the opinion that the decisions an individual makes on a daily basis can have substantial implications for their health (such as those related to diet, activity level, and rest). Wearable technology such as the Fitbit and the Apple Watch are capable of collecting vast amounts of data in real time from the sensors built into the devices. This type of information can shed light on a range of health-related habits that were previously unknown, such as how frequently one exercises or how well one sleeps, for example. The promise that wearables hold for application in medical research has not been able to live up to its full potential because self-measurement and electronic health records do not provide any correlation between the patient's state of health and the data collected by the wearable.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

*A. Literature Review*

For this literature research, we scoured the archives for information on Wearables data analysis techniques and practices.

An abundance of studies have been conducted thus far in the rapidly expanding field of wearable health and fitness monitors. Student health behaviors, particularly sleep and exercise habits, are analyzed statistically using Fitbit data.(Rachael, Stephen, Lixing, & Omar, 2016).

Evaluation of accuracy and security is required, as well as some familiarity with basic statistical methods.(Jason, Onyeka, Heather, & Gina, 2019).

Unsupervised testing of a certain statistical latent process model for joint segmentation of time-series data is framed as the problem of detecting temporal behavior from acceleration data. The most significant shortcoming of this method is that it makes use of unsupervised physiological signals in an explicit manner. (F, S, D, L, & Y, 2013).

Despite the fact that numerous researchers have utilized dimension reduction and outlier identification to identify abnormal sleeping habits in time series data, (Zilu, Mario, Chapa, & Takuichi, 2016) in addition to Hidden Markov Models for predicting alterations in heart rate (Sanghun, Chang-Sik, Sang-Ho, & Won-Seok, 2018) , If you're going to be employing the HMM model, you'll need to have a pretty good idea of how many hidden states there are and how many observations there are.

Many of the examined research also make use of data mining, machine learning, and deep learning. As an example, one study differentiates between stages of sleep by employing a deep neural hybrid network architecture composed of CNN and LSTM layers. (Shashank, Abdallah, & Pablo, 2019) .

In contrast, different research discussed the LSTM-Attention Ensemble approach, which demonstrated the highest accuracy in forecasting the following night's sleep efficiency and provided a more nuanced comprehension of the connections between sleep and activity.(Sungkyu, et al., 2019).

Characterizing the physical function PRO using a Random Forest classifier yielded an AUC of 76%, and an HMM was employed to find correlations. (Yiwen, et al., 2019). Heart disease has been classified using supervised classification algorithms, and the patient's vital signs have been discussed in relation to the condition.(Poojitha, Nikhita, Khurana, Sneha, & Weneen, 2019).

In (Zilu, Mario, & Chapa-Martell, Combining Resampling and Machine Learning to improve Sleep-Wake Detection of Fitbit Wristbands, 2019), For better sleep/awake prediction accuracy using Fitbit data, the authors have investigated the efficacy of predictive models combining the decision tree with resampling approaches.

K-means clustering is used to find measurement differences and categorize their dependence on the measurement level in (Yash, Jocelyn, Erich, & Steven, 2017).

1. OBJECTIVES
2. GATHER ALL STATISTICS OF THE DATASET
3. DATA CLEANING
4. DATA ANALYSIS
5. DATA PREPARATION
6. BUILDING REGRESSION MODEL
7. *Proposed Framework*

In our technique, the data exploration step is followed by the application of two regression models linear regression and random forest in order to make a prediction regarding the total amount of calories burned. The datasets that were utilized may be found in the repository that was provided by Kaggle. There are two CSV files, bringing the total number of cases to 15000, and each one has seven distinct features. The data set that was received from the Kaggle repository includes details about all of the participants, such as their gender, age, the length of their workout, their heart rate, their body temperature, their height, and their weight. This dataset is used to acquire the data that will be used for training. In addition to this, the second calories dataset has a target class that is made up of the calories burned by persons that are a match for each other.

1. *Preparation of Environment*
   * We are going use Google Collaboratory Environment for this research. It is very efficient at using scientific packages like numPy, pandas, matplolib, and sklearn..
2. *Data Cleaning and Manipulation*

Firstly, we observe and familiarize ourselves with the data. Next, check for the null values and impute those cells with the mean, median, or mode as per the column. Then we sanitize and scale the data so it will be easier to analyze patterns and findings.

In the dataset we dropped user id which was not useful in our prediction.

We created another feature from the existing dataset that is age groups we grouped the age in to three categories as shown in Fig1

As we expected, there is a significant difference between in counts of different age groups. Most of the people on this dataset are young. The second is middle-aged and the third one is old.

1. *Data Description*

# Analysis:

To perform z-score analysis, we analyze the number of rows in the dataset, the mean of the specific columns, and the standard deviation. minimum and maximum values in the data. We also analyze 25%, 50%, and 75% of the data.

We have the following features in our dataset: 1.User\_ID : The ID of the person which is unique. 2.Gender : Gender of the person.

1. Age : Age of the person.
2. Height : Height of the person in cm . 5.Weight : Weight of the person in kg .

6.Duration : Duration of the person's exercise/activity. 7.Heart\_Rate : Heart rate per min of the person.

8.Body\_Temp : Body temperature of the person in C∘ . 9.Calories : Calories burned in kilo calories.

See Fig 2 for more information.

# Visualizations from the Analysis

First, we find the frequency of use of these wearables across the week. How the users track their activity throughout the day How many times did they forget to use the device to log their activity? We use some scatter in a plot to visualize and find the type of correlation. What is the intensity of calories burned per hour and per step? We also analyzed some of the outliers due to some technical errors in the data. Finally, we are going to analyze the percentage of activity among the people. percentage of their activity and the relation between active time and sedentary time.

See Fig 3 for more information.

All of the aforementioned graphs show that the features of the dataset do not have a clear connection or link. There is no one-to-one correlation between, say, duration and either weight or height. This is because, regardless of BMI or height and weight, different people will have varying workout durations.

Features like duration and heart rate have a weak correlation with one another. Generally speaking (but with some uncertainty), it is reasonable to assume that a person's 'Heart Rate' per minute will increase in direct proportion to the amount of time spent exercising.

Height and weight, for example, have a strong relationship with one another.

The concept of Correlation can provide us with additional insights and advantages. However, this is all there is time for, as we shall delve deeper into the following chapter.

# Dependent and Independent Variables:

It's clear that the elderly has a higher basal metabolic rate and total energy expenditure compared to the younger and middle-aged populations. Interestingly, young people have the lowest calorie expenditure.

What's more intriguing is that females of all ages showed similar performance. That is to say, on average, they expended the same number of calories. In contrast, older men did better than their younger counterparts.

In addition, the early age group has a single outlier represented by a dot. The value of this point is larger than the sum of the third quartile and 1.5 times the interquartile range.

multiple traits are highly correlated with one another, we must choose one to keep and discard the others. The effectiveness of the model can be enhanced in this way. Our BMI column now includes both Weight and Height, which formerly appeared separately due to their substantial association as seen by the heatmap. The BMI can be preserved by eliminating the need for the Weight and Height columns.

1. *Learning Curve*

There is a clear peak at the training set size of 800 in both curves. It follows that the model's performance will not improve much if the training set size is increased above 800 examples. It turns out that we can get away with a training set size of only 800 instances without significantly impacting the model's performance. Even if more examples would have helped the training set and validation set achieve a lower MSE, the dataset does not contain enough meaningful features that our learning method can exploit to develop a more performant model.

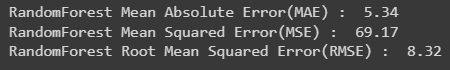
1. *Results*

# Linear Regression model

We have just implemented a simple Linear Regression to predict the calories burned with various parameters. Our RMSE for Linear Regression is about 12 which is acceptable. As just I said it is a simple model. We can reduce these errors by just replacing simple model with more complex model. Linear Regression Mean Absolute Error(MAE) : 8.52 Linear Regression Mean Squared Error(MSE) : 140.08 Linear Regression Root Mean Squared Error(RMSE) : 11.84 Random Forest Regressor Algorithm

We will be using the GridSearchCV on of the cross- validation methods that we use for selecting hyperparameters.

Specifically, the RandomForestRegressor method will have three hyperparameters: n estimators, max features, and max depth. Each of five possible splits will be run, and the one with the highest accuracy will be chosen.



we can see the RMSE for ReandomForestRegressor is lower than Linear Regression’s RMSE. It means that we can make better predictions with RandomForestRegressor.

See fig 5

1. CORREALTION

# Meaning of the results:

The heatmap displays the relationship between the two characteristics for each cell. It's clear that many different characteristics are highly correlated with one another. It's important to emphasize the need to eliminate as many extraneous components as feasible. The reason for this is that if we have a lot of features, the feature space dimension will be quite vast, and our model would run slowly on this feature. As a result, we'll have to eliminate some capabilities. If

The results will help find the answers to the research questions. We will be delivering our insights and providing recommendations based on our analysis. Here, we revisit our business questions and share with you our high-level business recommendations. The FitBit data set confirms that not all users fully utilize the functions of their devices or trackers. All 33 unique IDs used the step count function. 24/33 unique

IDs used the sleep tracking function. 14/33 unique IDs were used for heart-rate tracking. 8/33 unique IDs used their devices to track their weight. With the exception of the step count function, all other functions were used irregularly throughout the data tracking period.

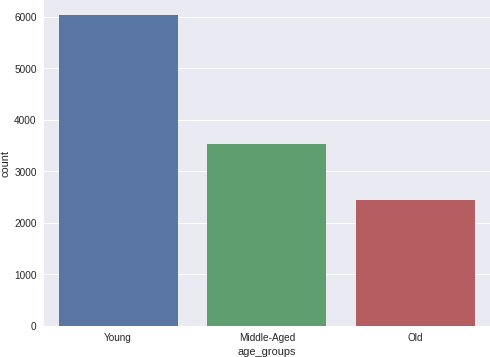
# Conclusion and Recommendation for Action

There is no one-to-one correlation between, say, duration and either weight or height. This is because, regardless of BMI or height and weight, different people will have varying workout durations.

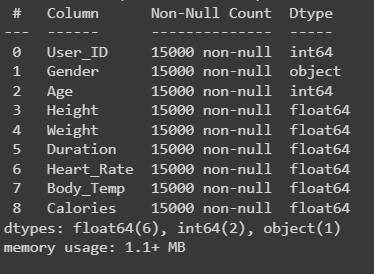
The vast majority of FitBit users (81.3% to be exact) are not utilizing the program for its intended purpose (monitoring healthy behaviors), but rather to keep tabs on how much sitting time they're getting in. Users are more likely to log their actions throughout the week than on weekends, possibly because they are more likely to be active during the week.

Both businesses create solutions aimed at helping women gain insight into their existing habits and make healthier choices based on that knowledge. The health and fitness industry is seeing a plethora of common trends that can easily be supplied to Bellabeat customers.

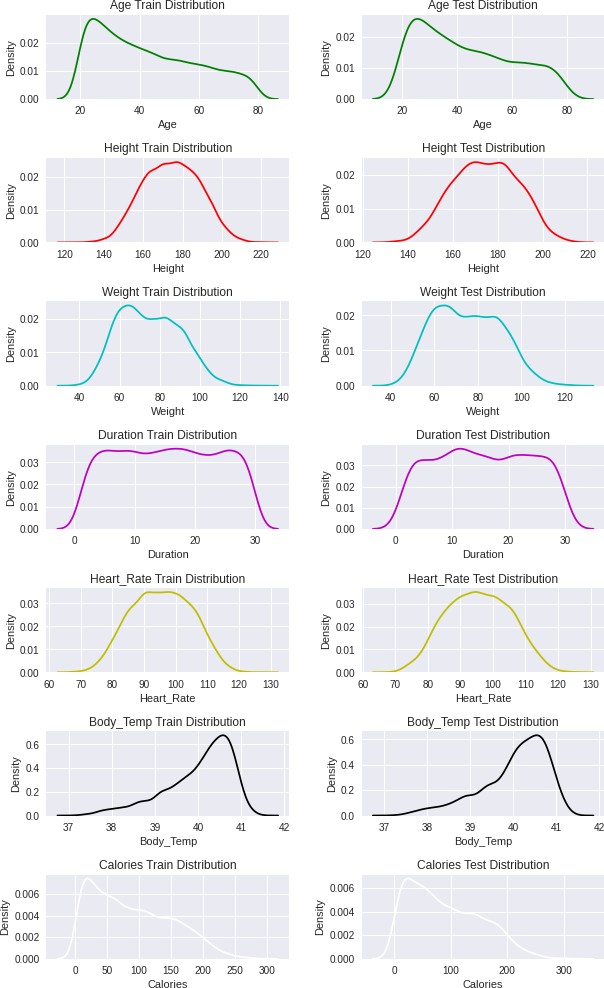
1. *Figures and Tables*
   1. Fig 1.



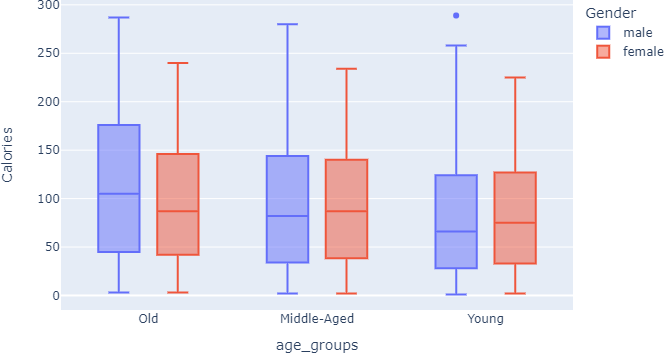
* 1. *Fig 2*



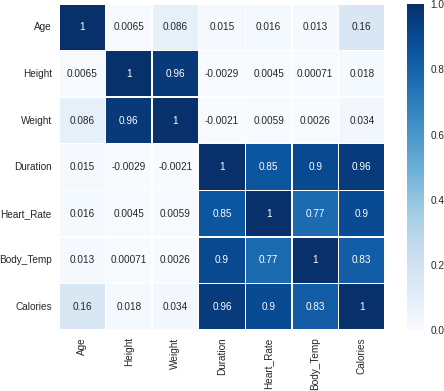
* 1. *Fig 3*



* 1. *Fig 4*



1. FIG 5



1. FIG 6

|  |  |
| --- | --- |
| [3] | P. Rachael, M. Stephen, S. Lixing and L. Omar, "Experiences Measuring Sleep and Physical Activity Patterns Across a Large College Cohort with Fitbits," ISWC 16, ACM, 2016. |
| [4] | O. Jason, E. Onyeka, Y. Heather and B. Gina, "A Look at the Security and Privacy of Fitbit as a Health Activity Tracker," ACM, pp. 241-244, 2019. |
| [5] | C. F, M. S, T. D, O. L and A. Y, "Joint segmentation of multivariate time series with hidden process regression for human activity recognition," Elsevier, 2013. |
| [6] | L. Zilu, A. Mario, M. Chapa and N. Takuichi, "A Personalized Approach for Detecting Unusual Sleep from Time Series Sleep-Tracking Data," IEEE International on healthcare informatics, 2016. |
| [7] | Y. Sanghun, S. Chang-Sik, L. Sang-Ho and K. Won- Seok, "“Forecasting of heart rate variability using wrist-worn heart rate monitor based on Hidden Markov Model," International Conference on Electronics, Information, and Communication, 2018. |
| [8] | R. Shashank, E. A. Abdallah and C. Pablo, "Poster: DeepSleep: A Ballistocardiographic Deep Learning Approach for Classifying Sleep Stages," ACM, 2019. |
| [9] | P. Sungkyu, Cheng-Te, S. H. Li, H. Cheng, W. L. Sang and C. Meeyoung, "“Learning Sleep Quality from Daily Logs," Conference on Knowledge Discovery and Data Mining ACM, 2019. |
| [10] | M. Yiwen, S. William, S. Chrisandra, J. Sandy, E. V.  E. Jennifer, N. B. M. C, L. Mayra, S. Brennan and W.  A. Corey, "A Machine Learning Approach to Classifying Self-Reported Health Status in a cohort of Patients with Heart Disease using Activity Tracker Data," IEEE Journal of Biomedical and Health Informatics, 2019. |
| [11] | A. Poojitha, R. A. Nikhita, S. Khurana, V. Sneha and  W. Weneen, "Personalized Health Monitoring Using Predictive Analytics," IEEE 5th International Conference on Big Data Computing Service and Applications, 2019. |
| [12] | L. Zilu, A. Mario and Chapa-Martell, "Combining Resampling and Machine Learning to improve Sleep- Wake Detection of Fitbit Wristbands," IEEE, 2019. |
| [13] | S. Yash, D. Jocelyn, H. Erich and L. Steven, "Wearables data integration: Data-driven modeling to adjust for differences in Jawbone and Fitbit estimations  of steps, calories, and resting heart-rate," Elsevier, 2017. |



# References

|  |  |
| --- | --- |
| [1] | C. Po-Han and D. Sujit, "Personalized Effect of Health Behavior on Blood," IEEE, 2018. |
| [2] | M. Yiwen, S. William, S. Chrisandra, J. Sandy and E.  V. E. Jennifer, IEEE Journal of Biomedical and Health Informatics, 2019. |

.