Jacob Son

VDS Coding Challenge Write-Up

Background Information

All my life, I have been on the skinnier side and was often made fun of by other peers and students. My parents also constantly told me to eat more, but due to my fast metabolism, eating large amounts of food had little to no effect on my weight. Upon coming to Vanderbilt, I made some changes to my diet and lifestyle: I started tracking my calories and other nutritional values on the Vanderbilt dining website, consistently going to the gym, and eating lots of protein to gain muscle and weight. Although my progress has not been very significant, these small steps that I am taking and the little improvements that I can visibly see have made me feel extremely fulfilled and accomplished.

As a result, I was interested in other people's diets, body mass indexes (BMI), and healthy lifestyles. I found a dataset with many subjects' BMIs, sex, step count, THR (target heart rate) fat burned, cardio levels, "Peak," age, and university class/year.

Dataset-https://figshare.com/articles/dataset/Dataset_The_Impact_of_Body_Mass_Index_on_Physical_Activity_and_Cardiac_Workload/13079396?file=25029731

Problem Statement

With this data, I wanted to cluster people's varying levels of health by analyzing their active lifestyles and statistics about their bodies. There will be 3 clusters: unhealthy lifestyles, moderately healthy lifestyles, and healthy lifestyles.

Hypothesis

Prior to the analysis, I expected that those with a higher THR, higher step count, higher fat burned, and higher cardio levels would be more healthy than those with lower values. Additionally, those with a BMI closest to 18.5 and 24.9 would be within the outer ranges of the healthy band, according to the BMI scale.

Methods

To first clean my dataset, I identified the number of null values, such as values that were missing or had values of 0. I found that my dataset had no null values.

Next, I modeled the THR values with 2 visual representations: a distribution graph and a box plot. From the distribution graph, it was clearly skewed left, meaning there were outliers in the higher range of the dataset. Similarly, from the box plot, there were multiple outliers far from the interquartile range in quartile 4, which showed the outliers in a different visual representation. To remove these outliers, I set upper limits and lower limits for the THR values by adding and subtracting triple the standard deviation from the mean THR. With these limits, I removed the THR values that were below the lower limit and above the upper limit. After the data was cleaned, 33 outliers from the THR values were removed.

I also used the same 2 visual representations to model the fat burned. Likewise, the distribution graph was skewed left, and the box plot contained outliers in quartile 4. Using the same method to calculate the limits and remove the values that don't lie in the limits, 22 outliers were removed from the fat burned.

Lastly, the class values were popped and removed from the dataset since the data provided age values, which are more accurate than students' year/class.

Once the dataset was cleaned, I used k-means clustering to cluster together the data, separating them by varying levels of healthy lifestyles. From the cleaned dataset, I removed all values except for the steps, THR, fat burned, and age. I set my k value to 3 to form 3 clusters: unhealthy, moderately healthy, and healthy lifestyles.

Results

From the three clusters, the unique colors/values were 1 (green/teal), 2 (yellow), and 0 (black).

The slightly green and teal cluster, cluster 1, had an average THR of 117.56, 110.39 fat burned, 5.53 cardio, and a peak of 1.63. These were the lowest values out of the 3 clusters, making this specific cluster consist of the least healthy individuals.

The black cluster, cluster 0, had an average THR of 122.38, 114.45 fat burned, 5.67 cardio, and a peak of 2.26. These moderate values out of the 3 clusters make cluster 2 consist of moderately healthy individuals.

The yellow cluster, cluster 2, had surprisingly high averages: an average THR of 284.18, 259.7 fat burned, 19.67 cardio, and a 4.8 peak. These values were significantly higher than the other clusters, so cluster 2 clearly consisted of the most healthy and active individuals.

For the evaluation metric, I used the silhouette score, which is used to calculate how distinguished and separate the clusters are from each other. With the 3 clusters, my score for the data was 0.4558. The silhouette score will be between -1 and 1, and a higher score indicates thata the clusters are more distinguished. My value was in the upper 72% of the silhouette range, which indicates that the clusters were relatively well defined and separate.

As a result, my initial expectations/hypothesis was correct; the higher one's heart rate, the more fat burned, and higher cardio levels indicate a more healthy and active lifestyle. The clusters shown in the graph aligned with the data; the yellow clusters, being the most significant and higher than the other clusters, were more isolated/separated than the black and green/teal clusters.

Resources

Dataset:

https://figshare.com/articles/dataset/Dataset_The_Impact_of_Body_Mass_Index_on_Physical_Activity_and_Cardiac_Workload/13079396?file=25029731