Lab 3: Healthcare Scenario - Healthy Living and Wellness Clustering

Amit Rajput

DeVos Graduate School, Northwood University

MGT 665: Solving Bus Problems W/ Machine Learning

Dr. Itauma Itauma

06/22/2025

Abstract

In this project, I was attempting to provide a health organization with greater insight into patient-level wellness by grouping people by their lifestyle habits. The dataset that I included consisted of exercise time in minutes, healthy meals per day, sleep hours, stress level, and bore an indication of body mass index (BMI). To separate similar patients I developed three clustering techniques: K-Means, Hierarchical Clustering and DBSCAN. I also utilized Principal Component Analysis (PCA) to reduce the overall complexity of the dataset and help visualize our clustering results. The silhouette scores were on the lower end but still provided useful clusters that can inform the development of personalized wellness programs.

In addition to evaluating the model outputs, I created labelled wellness profiles based on reference ranges for the various features. I organized the variables into high, medium and low categories to better interpret the cluster behaviours. In this phase, I noticed patterns such as low physical activity with high BMI or a combination of balanced things with 3 or 4 features. These wellness profiles can inform healthcare providers on the necessary baselines to capitalize on and develop actions that support long-term health improvements by consideration of the various groups of patients.

Introduction

Healthy living has become a primary focus in the current healthcare environment. For me, health upkeep means more than just steering clear of sickness. It means making lifestyle choices in a consistent manner such as exercise, eating healthy meals, managing stress, and getting proper rest. Many healthcare organizations are employing a wellness program that encourages healthy living, but it is challenging to implement a wellness program in consideration of an individual patient without using data.

With this project, I wanted to explore how data analytics and machine learning could segment patients into wellness profiles based on their habits. My view is that just because a health service provides for one group does not mean their program will be beneficial to every group. If health services would understand the group's specific needs, they could create and implement more effective and targeted wellness programming for each patient group. I decided to explore patient wellness data and use unsupervised learning methodologies to reveal behaviors and patterns.

Literature Review

Wellness is now recognized as much more than your physical health. It involves mental health, emotional health, social health and spiritual health (Myers et al. 2000). Researchers have identified lifestyle factors: Healthy Eating, Physical Activity, Stress Management and Sleep Preparation; that function to enhance wellness (Roscoe, 2009). These dimensions address modern health initiatives focused on the quest for a balanced life.

I found delineating patients into wellness clusters assists health professions in directing. Grumam et al. (20021) note clustering groups individuals with similar needs, especially for wellness and mental health programs and it makes sense to customize care based on behavioral data, rather than applying the same level of enhancement to every patient.

K-Means, or Hierarchical Clustering and other clustering algorithms has been widely employed in such research. These cluster techniques lend itself well in combination with Principal Component Analysis (PCA) involve the graphics and subsequently reveal the feature space structure that may not have been previously apparent to the naked eye (Keyes, 2002). While these methods do not always produce discrete data, they convey useful direction in developing effective wellness applications.

Methodology

Data Description & EDA

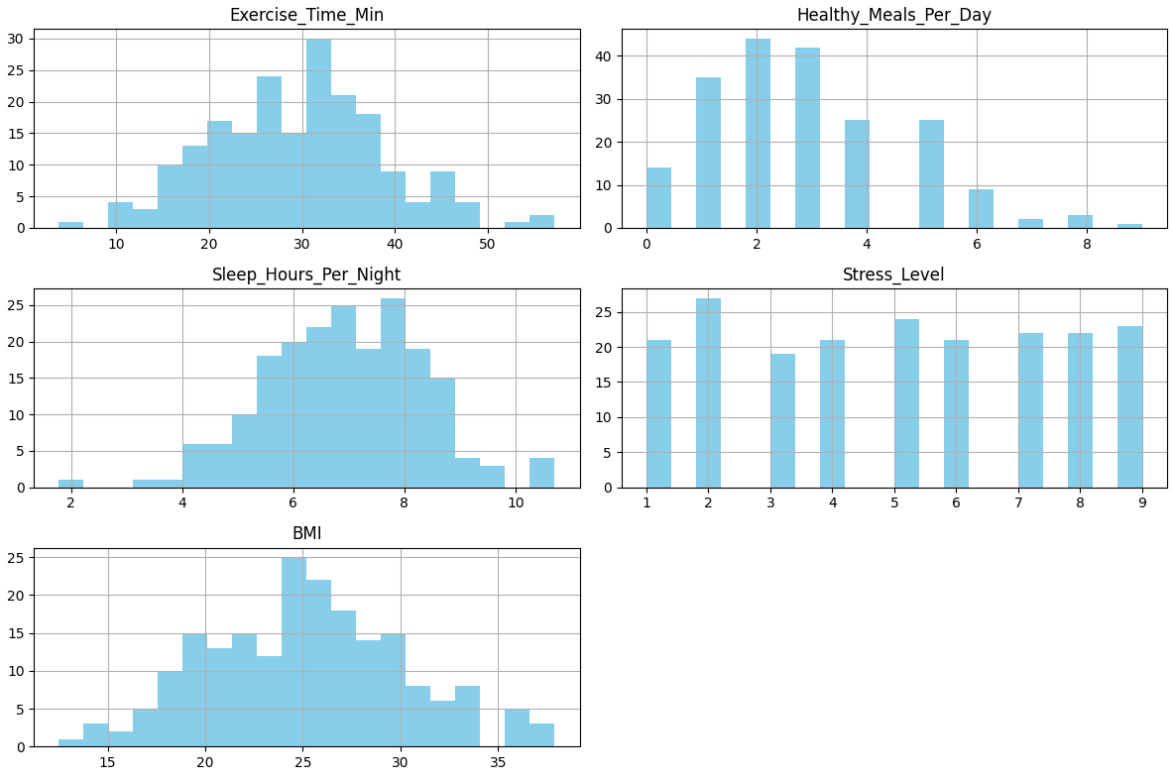
I started out by reviewing data from a dataset with health and wellness data for 200 subjects. The dataset included the following five features: exercise time in minutes each day, average number of healthy meals eaten per day, number of hours slept each night, stress level on a scale of 1 to 10, and BMI, or body mass index. To try and understand these variables better, I first looked at summary statistics and then constructed histograms for each feature. This allowed me to see central tendencies, spread, and distribution patterns.

Most patients in the dataset were completing 20 to 40 minutes of exercise each day. This suggests that, on average, across the group, these individuals were moderately active. In terms of dietary habits, most patients were consuming two to three healthy meals each day. In terms of sleep patterns, it appeared healthy in the dataset, with a lot of subjects sleeping six to eight hours each night. The stress ratings included a lot of variability, covering almost the full range of 1-10. This variability could be attributable to differences in lifestyle, occupation, or coping mechanisms. Most patients had BMI values between 20 and 30, generally considered normal to slightly overweight.

This exploratory data analysis (EDA) revealed a depth of understanding of these five features both independently and in regard to one another. For instance, I found some small correlations between BMI, time exercising, and stress. This preliminary analysis helped me build an overall intuition of the dataset, as the EDA provided context for my decisions around preprocessing and clustering in later steps of the project. Overall, the EDA was an important part of the entire wellness segmentation process.

Figure 1

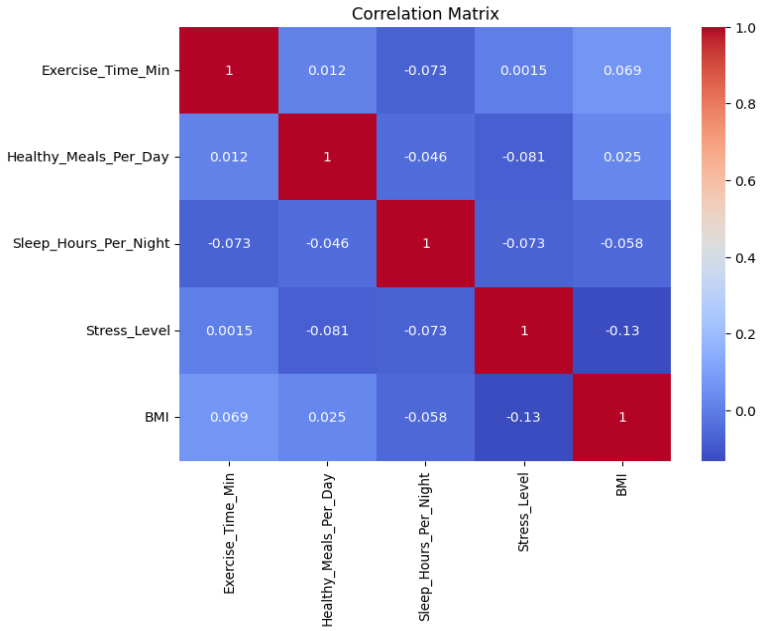
Distribution of Wellness Indicators



Note: This multi-plot figure shows histograms for five relevant health features: exercise time, number of healthy meals, sleep hours, stress level, and BMI. Most patients did exercise time for 20-40 minutes, had 2-3 healthy meals, and slept for 6-8 hours. The stress level and BMI had a wide distribution suggesting that there are various patterns of wellness within the population.

Figure 2

Correlation Heatmap



Note: The heatmap displays the pairwise correlations among the five health features. Most correlations are weak or very near zero, suggesting little interdependence between wellness indicators. The strongest negative correlation found is between stress level and BMI (-0.13), even this correlation is relatively small.

Standardization

To clean the data set for clustering, I standardized all variables using a standard scaler which ensured that each feature had equal weight and that clustering was not affected by any bias due to different scales. I then performed PCA to reduce the five features down to two principal components to make plotting and visually interpreting the clusters in two-dimensions easier while still retaining almost all of the variance from the data.

Table 1

Standardized Wellness Features for Sample Patients

	Exercise_Time_Min	Healthy_Meals_Per_Day	Sleep_Hours_Per_Night	Stress_Level	BMI
0	0.578767	1.173447	0.482957	-1.152351	1.565523
1	-0.104981	2.830078	-1.993156	0.771441	0.418669
2	0.741336	0.621237	-0.640956	-1.537110	-0.271010
3	1.683908	-1.035394	1.149993	1.156199	0.923359
4	-0.208235	0.069026	0.964166	-0.767593	1.146154

*Note***:** The values in this table are standardized values for the five health indicators after performing a scaling operation on the dataset. Positive values indicate above-average wellness behaviors, while negative values indicate below-average wellness behaviors. Standardizing the variables ensures that all variables contribute equally during clustering.

Clustering Techniques

I attempted three different clustering methods to segment the patients on the basis of their wellness behaviors. K-Means, Hierarchical Clustering, and DBSCAN. For K-Means, I tried many values for k and decided on k = 7 clusters. The elbow method led to that choice, but also the silhouette scores, though not very high, revealed that k = 7 created the most meaningful segmentation since the patients were able to be grouped in more relevant segments of wellness behavior. Next, I performed hierarchical clustering and created a dendrogram to use as a basis for determining where to cut the tree. I decided on four clusters, which put the patients into broader categories of wellness. Finally, I tried DBSCAN clustering which is a density-based clustering algorithm that only seems relevant in some datasets. With those two conditions, it shouldn't be a surprise DBSCAN didn't work well on my data. In the end, it simply found a major cluster of patients while identifying so many other points as noise. The major cluster was pretty straightforward since patients showed a high level and similarity in the features and had little overlap in feature values. This overlap,

however, contributed to the lack of clearly defined dense areas which would have made DBSCAN appropriate for this analysis or technique.

Evaluation Metrics PCA

To simplify the dataset and facilitate the visualization process, I completed Principal Component Analysis (PCA) after standardizing the data. PCA enabled me to convert the original five observations into two principal components, which retained most of the relevant information, while allowing me to plot and interpret the clustering results in a two-dimensional space. For measuring the quality of clustering, I relied on silhouette scores and Within-Cluster Sum of Squares (WCSS). Silhouette scores measure how well each observation conforms to the clustering structure (i.e., the score indicates better separation with higher scores). WCSS measures the compactness of the clusters; higher values of WCSS indicate the clusters are more separated, whereas lower values of WCSS indicate more compact clusters. By using these metrics, I could compare the relative performances of the various clustering methods in terms of how well their clustering processes articulate and capture the patterns in the data.

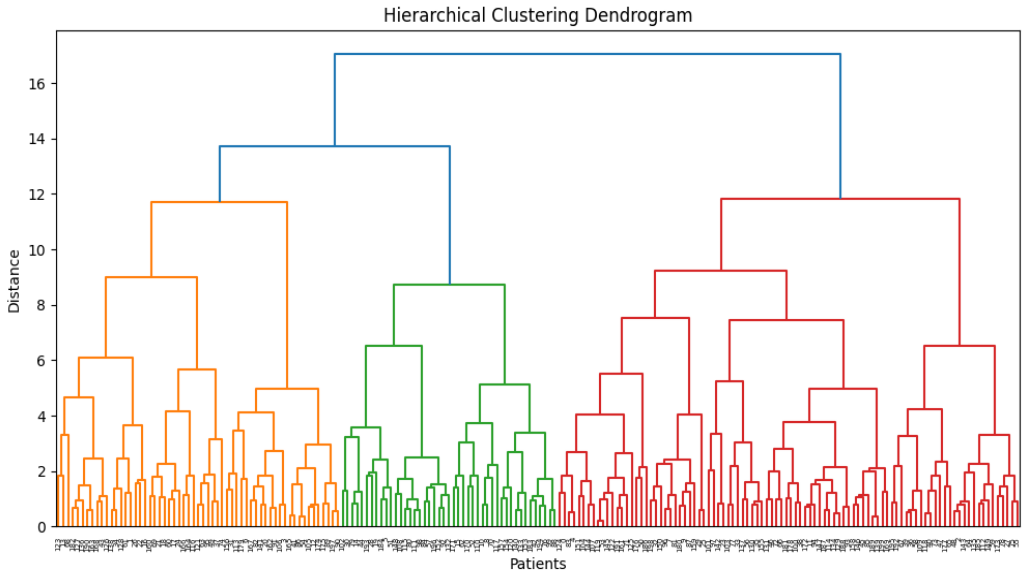
Results

Cluster Findings

Each algorithm produced a different grouping of patients. K-Means formed seven clusters, showing unique combinations of wellness features such as high stress with low exercise or balanced habits across multiple features. Hierarchical Clustering produced four broader groups with less separation. DBSCAN couldn’t form meaningful groups, possibly because the data lacked high-density areas.

Figure 3

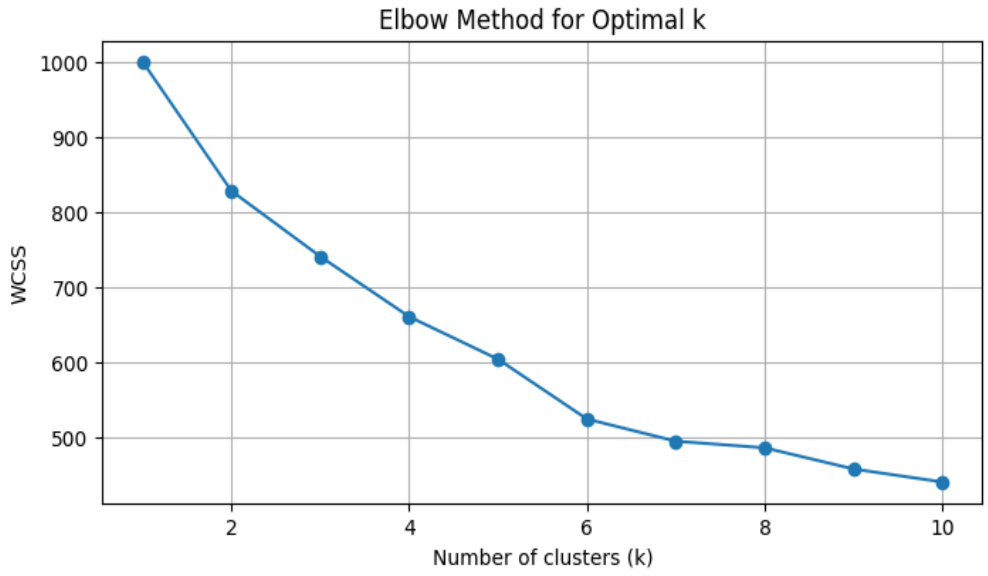
Hierarchical Clustering Dendrogram



Note: The dendrogram shows the outcomes of hierarchical clustering, where patients are grouped based on how alike they are; vertical lines indicate mergers between clusters, height indicates distance (or dissimilarity). By cutting this tree at an appropriate height, we arrived at four unique clusters going forward.

Figure 4

Elbow Method for Optimal k



Note: This graph shows the value of the Within Cluster Sum of Squares (WCSS), for different values of k in K-Means clustering. The "elbow point", (where the WCSS declined, then very much slowed in its decline), is somewhere around k=7 in the graph.7, The optimum value of k occurs at this elbow point, finding a balance between complexity of the clustering model and its performance.

Silhouette Scores and WSS

To evaluate clustering quality, I used the silhouette score:

- K-Means (k=7): Silhouette Score - 0.178
- Hierarchical (k=4): Silhouette Score - 0.114
- DBSCAN: Not applicable

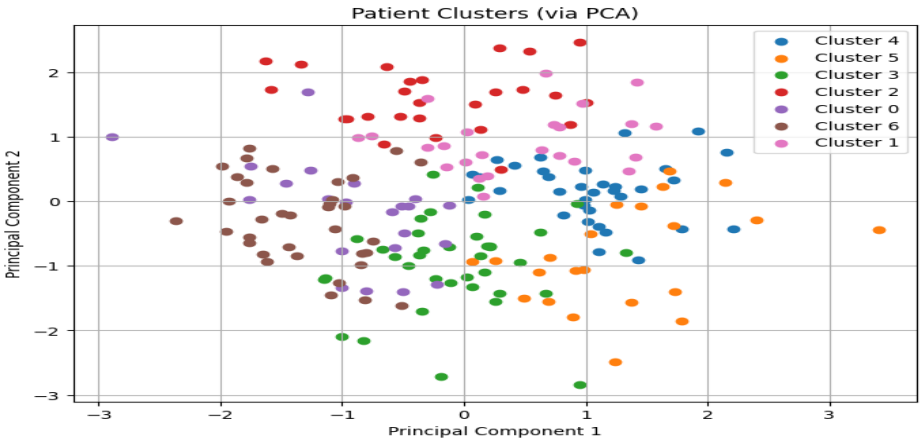
The scores were low, indicating weak separation between clusters, but exhibit some underlying structure. I calculated the Within-Cluster Sum of Squares (WCSS) for K-Means, which bolstered my case that k=7 was the right amount of clusters.

PCA Interpretation

PCA provided the dataset with two principal components, with the first being variation associated more with BMI and stress, followed by the second relative to sleep and exercise. I visualized the clusterings with PCA, and while the ability to visually separate clusters was not that clear cut, I could still see areas of separation for some of the clusters.

Figure 5

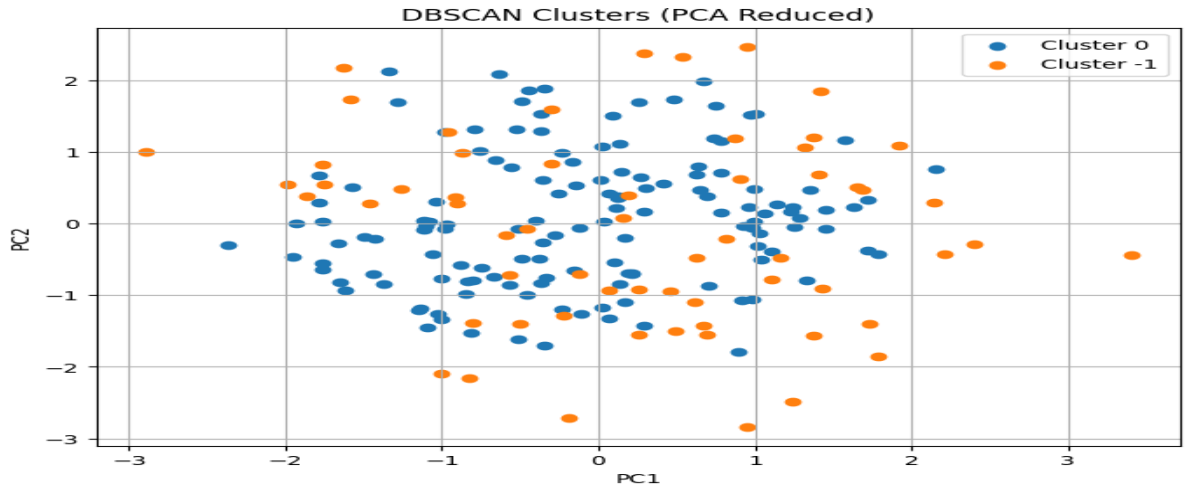
Patient Clusters (via PCA)



Note: The scatters plot displays patient clusters assigned to each patient through K-Means, which have been transformed into two dimensions with Principal Component Analysis (PCA). The clusters will also display different colors. PCA will take the five original features and reduce them to two principal components. The clusters have some overlap, but you can see some of the clusters have separation and can represent different wellness.

Figure 6

DBSCAN (After PCA)



Note: This scatter plot illustrates the clustering results from DBSCAN after we applied PCA to perform dimensionality reduction. Almost all the data points were assigned to one cluster

(Cluster 0) while the remaining points were assigned to noise (Cluster -1). There are no clearly separated groups, suggesting that DBSCAN was not able to pick up on distinct clusters that were embedded in this dataset.

Discussion

Interpretation of Results

Even though my clusters weren't perfectly separable, they still informed me of a few things. I was able to ascertain clusters like one group of patients who exhibited low activity and high stress or another group who exhibited better habits across all features. The patterns I have shared are useful for healthcare practitioners who will want to adopt specific personalized interventions.

Limitations

The other limitation I encountered was the low silhouette scores. These scores highlighted that the clusters were overlapping and not distinctly separate from one another. I believe this happened because of the limited number of features and the inevitable variability in patient behaviours. DBSCAN did not work either, because there were no very dense clusters present in the dataset.

Insights for Wellness Intervention Design

With these limitations in mind, the clustering analysis was still helpful in providing a launch Pad for personalized healthcare. For example, one cluster may benefit from sleep intervention, while others may need dietary and exercise intervention. If health organizations develop wellness plans based on the clusters that are identified, they can ultimately provide patients with more meaningful support.

Conclusion

The project enabled me to investigate how machine learning techniques could be used to identify wellness profiles using lifestyle behavior data from patients. I was able to use clustering techniques such as K-Means, Hierarchical Clustering, and DBSCAN to better define segments that emerged within the dataset of five-related wellness indicators—exercise time, healthy meals, sleep amount, stress level, and BMI. While the clusters weren't distinctly defined, we were able to identify some meaningful wellness profile groups that could be valuable for a healthcare provider to offer tailored interventions. The analysis ultimately included principal component analysis (PCA) to aid in the interpretation of the dataset, which helped to simplify and visualize the clusters. Overall, the study was interesting and demonstrated the ability to realistically use unsupervised learning to help facilitate data-informed planning of wellness programs.

In light of what I discovered, I would recommend adding features like age, gender, and pre-existing medical factors to better capture clusters. I also recommend acquiring data across time, and watching for trends to understand the evolution of wellness behaviors. These recommended steps can allow practitioners to intervene at a more precise level to develop personalized wellness interventions for specific patient groups. Moving forward, I will also evaluate supervised learning models to predict future health outcomes and explore the potential of wearable device data for real-time wellness monitoring. I am optimistic that as we build richer datasets and sophisticated analytical methods, we will be able to offer even finer and more useful wellness profiles for preventive health responses.

References

Gruman, J. A., et al. (2021). *Well-being and mental wellness*. Retrieved from https://www.researchgate.net/publication/351782904_Well-Being_and_Mental_Wellness_Well-Being_and_Mental_Wellness

Keyes, C. L. M. (2002). *The mental health continuum: From languishing to flourishing in life*. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC5508938/

Myers, J. E., Sweeney, T. J., & Witmer, J. M. (2000). *The Wheel of Wellness: Counseling for wellness: A holistic model for treatment planning*. Retrieved from https://www.tandfonline.com/doi/full/10.1080/21642850.2021.2008940

Roscoe, L. J. (2009). *Health to wellness: A review of wellness models and transitioning back to health*. Retrieved from https://www.researchgate.net/publication/329258077