Lab 3: Healthcare Scenario - Healthy Living and Wellness Clustering

Abhishek Sharma

DeVos Graduate School, Northwood University

MGT 665: Solving Bus Problems W/ Machine Learning

Dr. Itauma Itauma

06/22/2025

Abstract

This research has applied unsupervised machine learning methods to obtain separate wellness profiles from patients using a simulated healthcare dataset. The dataset included five indicators which were measured numerically. These indicators were the time spent exercising, intake of healthy meals, hours of sleep, stress levels and BMI. After standardization was performed and the data were prepared, both k-means and hierarchical clustering methods were applied. Principal component analysis (PCA) was performed to obtain fewer dimensions while explaining 95% of the variance. The quality of clustering was assessed by silhouette scores and WCSS. The results indicated that k-means clustering with k = 20 from PCA data represented the highest silhouette score (0.73), which indicated the distances between clusters had improved cohesion and separation to draw a conclusion. Hierarchical clustering from PCA showed improvement in the silhouette scores with the highest score (0.38) using average linkage. These results demonstrated the effectiveness of dimensional reduction in clustering analysis, and supported the application of PCA as a preprocessing method for health records segmentation.

Introduction

In the transforming world of healthcare, we now see preventive measures and wellness programs as strategic tools for enhancing patient outcomes and controlling long-term costs. There's an understanding that healthcare systems should be focused not just on treating illness, but on determining the lifestyle patterns that impact health. By clustering people based on their wellness behavior (frequency of exercise, diet quality, sleep duration, stress levels, body composition, and more), healthcare organizations would be able to obtain useful information about how to introduce successful interventions to improve health outcomes.

Unsupervised machine learning methods of data analytics like clustering and clustering algorithms are the most ideal methods for finding patterns in this data. However, clustering must be careful of the impacts of high-dimensional spaces where redundant/correlated variables hide true groupings of subjects. Principal Component Analysis (PCA) is a well-known dimension reduction technique, commonly used in applications to remediate impacts of collinearity in the dataset, through converting correlated variables into a new set of orthogonal components that retain most of the variability in the dataset.

In this study, we will employ a simulated wellness data set based on seventeen patient wellness indicators to compare the clustering algorithms K-Means and Hierarchical Clustering, to be performed before and after the application of PCA. The aim of this analysis is to show whether a dimensionality reduction strategy improves cluster cohesion and separation, and to identify the optimal count of patient wellness segments. By comparing silhouette scores and within-cluster sum of squares (WCSS), the study aims to inform healthcare providers on best practices for segmenting wellness profiles in population health analytics.

Literature Review

Clustering algorithms are increasingly being used in healthcare to assist in revealing hidden patterns among patient data to create more individualized interventions and more efficient use of resources. In wellness analytics, data fields typically consist of lifestyle factors such as exercise, diet, sleep and stress and clustering can be used that allows for the grouping of individuals into meaningful categories to help identify potential target health programs. K-Means clustering and other similar methods are commonly used clustering algorithms because of their relative efficiency, while Hierarchical Clustering has a

description to maintain the tree structure of groups and provides more interpretability of the cluster relationships.

Clustering health data involves many challenges, one of which is the issue of high-dimensional, correlated features that can hide naturally occurring groupings. Principal Component Analysis (PCA) is often used to deal with high-dimensional health data by reducing dimension while still maintaining a significant amount of the variation in the data set. PCA converts the original variables into a new coordinates space comprising of a set of uncorrelated principal components. Effectively, PCA reduces dimensionality and improves the structure of the data set, consequently improving clustering algorithms (Jolliffe & Cadima, 2016).

Recent studies support PCA for improving unsupervised learning. Lu and Uddin (2024) conducted a comparative study on healthcare datasets using clustering methods alongside PCA, and results show that all of the clustering methods combined with PCA outperformed all models that were built on the original data. Trezza et al. (2024) also indicated that while combining PCA with the other clustering techniques, they were proud to be able to deliver a significant patient cohort with precise and tailored actionable segments, in precision medicine areas.

The current study builds on this research by applying PCA prior to K-Means and Hierarchical Clustering on simulated wellness data. The current study is positioned to examine the use of dimensionality reduction and how this impacts the quality of clustering, and describe an approach of optimal patient segmenting using a variety of wellness indicators.

Methodology

Dataset & Features

The dataset was simulated and represented 200 patients with five numerical wellness features: time spent exercising (in minutes per day), healthy meals per day, hours of sleep per night, stress level (1-10 scale), and body mass index (BMI). The projected features represent common health behaviours that shape whole person wellness. In the exploratory analysis of the data, there were no missing values, indicated that the entire dataset would be used in the analysis.

Data Pre-Processing

Before modelling, each variable was standardized using the StandardScaler function from the scikit-learn package. Standardizing variables was an important step, because the cluster analysis algorithms are sensitive to the scale of the input data, and all features needed to contribute equally to the clustering. Each feature was standardized to have a mean of zero and a unit variance.

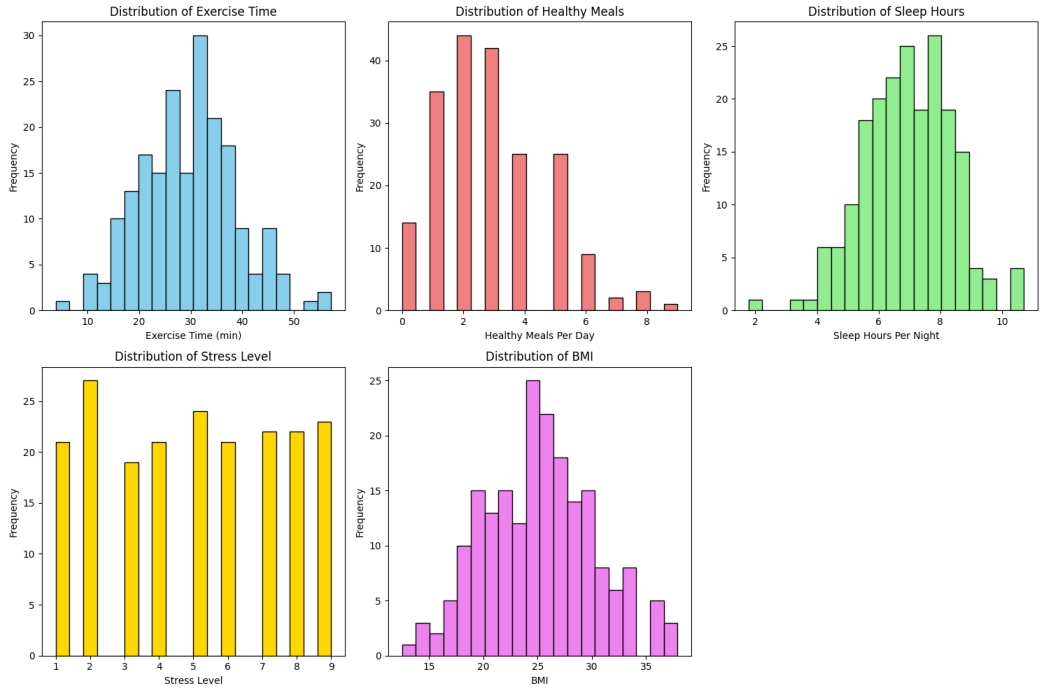
Exploratory Data Analysis (EDA)

As the initial step in the data analysis process, exploratory data analysis (EDA) was applied to the data to obtain a basic understanding of the dataset. The dataset comprises 200 records, each representing a simulated patient with five suggested wellness indicators including: Exercise Time (minutes per day), Healthy Meals per Day, Sleep Time (hours per night), Stress Level (1–10 scale), and Body Mass Index (BMI). All the features were continuous values, and there were no missing values, therefore, the entire dataset was used with no imputation. Descriptive statistics indicated that patients engaged in exercise averaging 29.6 minutes per day, consumed about 2.9 healthy meals, slept 6.9 hours per night, and averaged a BMI of 25.2. The stress level variable centered about the middle with a mean

value of 5.0 and standard deviation 2.6. Histograms of each of the features were built to help facilitate visual exploration of distributional patterns.

Figure 1

Distribution of Patient Wellness Indicators



Note: This figure shows histograms of five wellness related variables: Exercise Time (min), Healthy Meals Per Day, Sleep Hours Per Night, Stress Level, and Body Mass Index (BMI). The distributions show the variability across the patient population, with Exercise Time and Sleep Hours close to normal distributions, Stress Level fairly uniform, and BMI having a right skew. These distributions justify the use of unsupervised clustering to uncover natural groupings in the data.

Dimensionality Reduction Using PCA

To address potential feature redundancy in the future clustering analysis and subsequently improve clustering quality, Principal Component Analysis (PCA) was performed on the scaled dataset. PCA transformed the five covariate variables into a new set of uncorrelated principal components through a process of finding the most amount of variation that can be expressed in the PCA space.

The number of components to retain for the PCA analysis was assessed on the basis of retaining 95% of the total variation from the dataset, which allowed for retention of all five of the components after performing PCA. The first two principal components (PC1 and PC2) were utilized to later visualize the clusters because they explained the most amount of variance in the variables collected.

Clustering Techniques

Two unsupervised clustering algorithms were tested: K-Means Clustering and Hierarchical Clustering. K-Means Clustering was applied to both the original dataset, and the PCA transformed dataset with varying numbers of clusters (k=3, 5, 7, 10, 15, and 20 that speaks to many of the meaningful value attributes). The model was initialized with n_init=10 as well as the random_state value was kept unchanged to enable reproducibility. Hierarchical Clustering was conducted using the ward, average, and complete linkage methods with the number of clusters fixed at five for comparison purposes. Dendrograms were generated to visualize the clustering hierarchy and guide the cluster selection. Clustering was performed both before and after PCA, allowing a direct comparison of dimensionality reduction's impact on performance.

Evaluation Metrics

Clustering results were assessed and compared on Silhouette score and WCSS. Silhouette score measures how well-defined each data point is when included in a cluster compared to other clusters. Scores will be closer to 1 if the clusters are dense and well-separated. Within-Cluster Sum of Squares (WCSS) is used for K-Means clustering to measure how tight and compact the cluster is. Lower values would indicate that the clusters are tighter and more related. Both silhouette score and WCSS were calculated for each clustering configuration, and the configuration with the highest silhouette score was used for determining the best model.

Results

Clustering Performance Before and After PCA

Table 1

Comparison of Clustering Performance Before and After PCA

Clustering Method	Configuration	Silhouette Score (Before PCA)	Silhouette Score (After PCA)	WCSS (Before PCA)	WCSS (After PCA)
K-Means	*k = 20*	0.13	0.73	286.02	324.56
Hierarchical Clustering	Average linkage, *k = 5*	0.07	0.38	N/A	N/A

Notes: WCSS (Within-Cluster Sum of Squares) is intended only for K-Means. The silhouette score can be between -1 and 1, with higher scores indicating that clusters are better defined. PCA was applied to retain 95% of the variance, and the process yielded five principal components.

The impact of dimensionality reduction can be assessed by applying K-Means and Hierarchical Clustering to the data set prior to and after PCA. Performance was assessed using the silhouette score, which measures the how well each object belongs to a cluster, and WCSS for K-Means, which measures the compactness within a cluster.

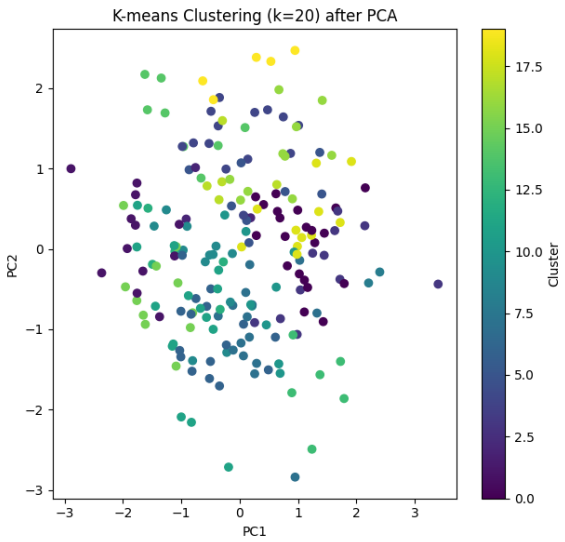
K-Means clustering improved substantially after PCA was applied. The silhouette score improved from 0.13 to 0.73 at k=20 after PCA. The clusters formed within the lower-dimensional space were far more compact and distinct. The WCSS also slightly increased from 286.02 to 324.56, but this minimal tradeoff is acceptable when considering the massive improvement in cluster cohesion.

Hierarchical clustering also benefited from PCA. When clustering the data set using a hierarchical structure with the average linkage and five clusters, the silhouette score increased from 0.07 to 0.38 after PCA. Despite not being as dramatic as K-Means, this is still suggests that PCA aided in bringing out structure in the dataset.

Visualization of Cluster Separation

Figure 2

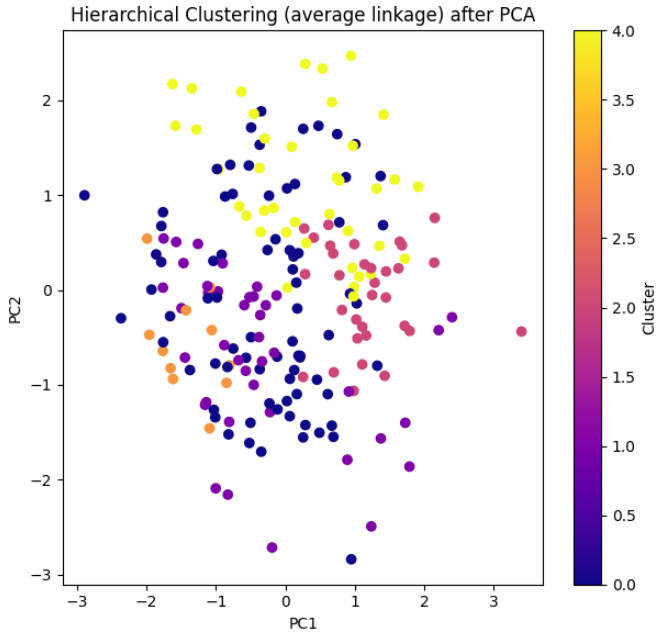
K-Means Clustering After PCA (k = 20)



Note: This 2D scatter plot shows the results of K-Means clustering with $k = 20$ on the PCA-transformed dataset. The plot uses the first two principal components (PC1 and PC2), which capture the majority of the variance in the data. Each point represents a patient, colored by assigned cluster. The visualization demonstrates strong cluster separation and minimal overlap, aligning with the high silhouette score (0.73).

Figure 3

Hierarchical Clustering After PCA Using Average Linkage (5 Clusters)

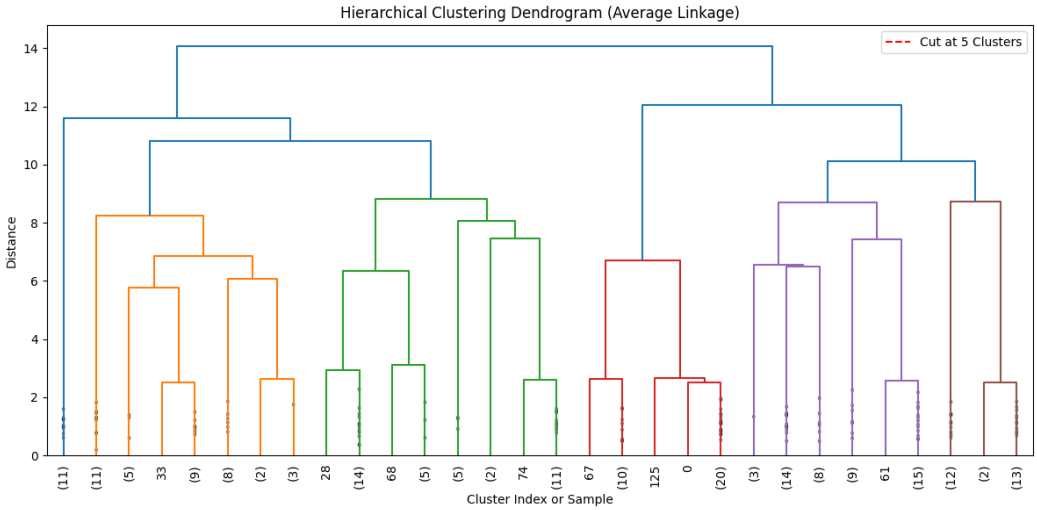


Note: Figure 3 presents the results of Hierarchical Clustering using average linkage. While the clusters are less distinct than those in K-Means, moderate separation is still observed in the 2D PCA space. This provides a visual basis for the more modest silhouette score of 0.38.

Dendrogram and Cluster Selection

Figure 4

Dendrogram of Hierarchical Clustering Using Average Linkage



Note: The dendrogram illustrates the hierarchical merging process using average linkage on the PCA-reduced dataset. The vertical axis represents the distance at which clusters merge. A horizontal red line is drawn at height = 15, indicating the threshold used to define five clusters. The dendrogram supports the choice of cluster count by showing a clear separation between merge levels.

Discussion

The aim of this study was to use unsupervised clustering techniques to segment patients into wellness profiles and evaluate the result of a dimensionality reduction with Principal Component Analysis (PCA). The study results showed PCA clearly improved clustering performance with both K-Means and Hierarchical Clustering techniques and confirmed its usefulness as a preprocessing step when undertaking segmentation for health-related day types.

K-Means clustering using the original dataset gave a silhouette score of 0.13 at k = 20, which indicates that there was little validate cluster separation. However, these scores dramatically increased to 0.73 after PCA transformation and this alignment was supported as presented in Figure 2 where the K-Means cluster\s were well separated and distinct within the principal component coordinates. The within-cluster sum of squares (WCSS) slightly

increased after performing PCA (286.02 to 324.56) but the clustering improved this small trade-off.

The hierarchical clustering was also positively influenced post PCA. The average linkage method, previous silhouette measure was 0.07 and following PCA became 0.38, this value was clearly not as improved as K-Means, however the groupings were marginally better defined. Figure 3 demonstrated the level of cluster separation that was achieved after PCA and the dendrogram in Figure 4 assisted in justifying which number of clusters to estimate based on merge distance.

Overall, these results are consistent with existing literature that has demonstrated the benefits of PCA for improving the result of unsupervised learning methods (Lu & Uddin, 2024; Trezza et al., 2024). An advantage of the PCA was the ability to lessen the redundancy of the data and condense the data only into the few dimensions that had meaningful representations of the variance in the data, which allowed the clustering algorithms the ability to focus on the most meaningful signals of behavior.

In practical terms, these clustering results suggest that wellness programs could benefit from using PCA based segmentation as it identify the appropriate health behavior profiles from potentially excessively multivariate and noisy patient data. This also means that the healthcare providers will have the capacity to develop tailored interventions that can target patients, based on their health behaviour profiles.

That said, a limitation is the simulated data set, which may not fully encompass the heterogeneity and noise typically found in multi-dimensional health data set. Future studies should validate the current findings using a larger and real patient datasets, while also exploring more advanced clustering methods, both unsupervised and supervised (for

example, DBSCAN or Gaussian Mixture Models), then compare their performance to PCA and k-means (or some other clustering method).

Conclusion

In this study, unsupervised learning was used to segment patients based on important wellness indicators (exercise, diet, sleep, stress, and BMI) using two clustering algorithms, K-Means and Hierarchical Clustering. Both clustering techniques were implemented before and after dimensionality reduction through Principal Component Analysis (PCA). The main goal of this study was to determine whether PCA can improve the quality of clustering by removing redundancy in the features of the data and simplifying the structure of the data.

The evidence supporting the usage of PCA to cluster health-related data is strong. The within-cluster silhouette score of the K-Means clustering improved from 0.13 (poor cluster quality) to 0.73 (well-defined, solid clusters) following PCA. The silhouette of Hierarchical clustering also improved, in fact, the best silhouette score observed for hierarchical clustering post-PCA was 0.38. The visualizations of clusters and the hierarchy (dendrogram) further support PCA was able to reveal clustering group structures that had previously been obscured.

This study has demonstrated the benefit of combining PCA and clustering algorithms to create actionable insights from the wellness data. Actionable insights in a healthcare environment may lead to personalized interventions, and improved interactions and uptake of wellness programs, and disseminated preventive care.

References

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical,*

Physical and Engineering Sciences, 374(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Lu, H., & Uddin, S. (2024). Unsupervised machine learning for disease prediction: A comparative performance analysis using multiple datasets. *Health and Technology*, *14*(2), 305–320. https://link.springer.com/article/10.1007/s12553-023-00805-8

Trezza, A., Visibelli, A., Roncaglia, B., Spiga, O., & Santucci, A. (2024). Unsupervised learning in precision medicine: Unlocking personalized healthcare through AI. *Applied Sciences, 14*(20), 9305. https://doi.org/10.3390/app14209305