**Srinivas Bhootam**

**DeVos Graduate School, Northwood University**

**115537-MGT-665-NW Solv Probs W/ Machine Learning**

**Unsupervised Segmentation of Patient Wellness Profiles for Targeted Health Interventions**

**Dr. Itauma Itauma**

**June 22nd, 2025.**

**Unsupervised Segmentation of Patient Wellness Profiles for Targeted Health Interventions**

**Abstract:**

This paper demonstrates that we can sort individuals by their health habits and design more effective wellness programs. We created a fake dataset of 200 adults that featured minutes of exercise, healthy meals eaten, hours of sleep, stress level and BMI. We first performed the two cluster analyses (K-Means and hierarchical) on the selected five variables. Next, we performed principal component analysis (PCA) to reduce the data into a two-dimensional space where we could easily plot it. We then measured how well they held together (silhouette score); K-means on 5 dimensions got 0.168, hierarchical got 0.114, and K-means on the two dimensions from PCA got 0.152. Ultimately, we discovered four charmed clusters, each with their own combination of habits. These segments can inform targeted interventions, such as stress workshops or peer mentoring. Simple unsupervised learning and PCA can be effective for designing targeted, data-driven health programming, as we demonstrate in this study.

**Introduction:**

Diseases such as hypertension, diabetes, and obesity associated with lifestyle factors are a significant burden on health care systems worldwide. Channeled preventive policies can help to reduce risks related to physical inactivity, bad nutrition, lack of sleep and stress. Unsupervised learning termed cluster analysis—permits the identification of natural groupings of patients for the optimization of tailored wellness interventions (Kaufman & Rousseeuw, 2009). Yet interpretation and communication of high-dimensional health data are challenging. PCA aims to address this by projecting data onto orthogonal axes that explain the most variance, so visual interpretation is enhanced without overly compromising cluster quality (Jolliffe, 2011).

**Literature review:**

Clustering is a crucial method in the application of statistical learning in healthcare to discover sub-populations of patients for intervention. K-Means reduces intra-cluster variance by assigning patients to a centroid closest to the patient, which generates compact sets, whereas HCA constructs a tree of nested clusters that can be pruned at various levels (Kaufman & Rousseeuw, 2009). The silhouette coefficient assesses the extent to which each patient belongs to its cluster relative to the nearest neighboring cluster, higher being better discrimination (Rousseeuw, 1987). PCA used together with clustering lowers dimension factors, retains kernel information, decreases noise and may facilitate more faithful visualizations, and entails only a small loss of cluster quality (Jolliffe, 2011). This hybrid approach has been successful in-patient stratification in management of chronic disease and has permitted focused resource allocations including targeted nutrition or stress management programs.

**Methodology:**

**Description of the Data and Preprocessing**

We simulated a cohort of 200 adults who each had five continuous measures of health: minutes of exercise per day, number of healthy meals per day, hours of sleep per night, stress level on a 0–10 scale, and body mass index (BMI). After eliminating any redundancy and system errors, the rows with missing values, which were less than 2%, were deleted. To have all variables on an equal footing, and avoid any one variable dominating the analysis, we standardized each to have zero meaning and unit variance.

**Clustering Procedures**

**K-Means Clustering**

We implemented Lloyd’s K-Means algorithm, which initializes centroids using k-means++ method to speed up convergence and achieve stability. Using the elbow method to plot within-cluster sum of squares (WCSS) for k between 2 and 10, we observed a sharp elbow at k = 4, which we chose. We used K-Means due to its computational speed and its natural ability to discover tight, more-or-less spherical clusters in first-stage wellness data.

**Ward’s Linkage Hierarchical Clustering**

We used Ward’s linkage criterion, which joins clusters in a way that minimizes the total within-cluster variance, to construct a dendrogram and again selected four clusters at a cut-off that balanced interpretability versus cohesion in cluster membership. Ward’s method was used because it generates relatively equal-sized clusters that are not sensitive to noise in small datasets, which is appropriate for exploratory segmentation of health measures.

**Dimensionality Reduction**

We scaled the 5 extracted features and subjected them to principal component analysis (PCA), which resulted in the first two principal components (PC1 23.7% of variance and PC2 22.1% ) that explained 45.8% of variability. PCA was added to reduce data dimensions, minimize multicollinearity between highly correlated wellness’ indicators, and create a two-dimensional representation which makes the cluster structure easy to interpret visually.

**Evaluation Metrics:**

In order to measure the quality of each clustering solution, The silhouette coefficient for:

* Original 5-dimensional data: K-Means
* Ward doesn't make any assumptions on the original data
* K-means on the 2d PCA full-reduced data

Silhouette score (between -1 and +1) for each sample indicates how the sample is closely associated with its own cluster and ought to be separated together with the neighboring cluster and can be used for comparison of clustering objects.

**Results**

A diagram of a well-known wellness indicator

AI-generated content may be incorrect.

A graph of a number of blue bars

AI-generated content may be incorrect.A graph of a number of meals per day

AI-generated content may be incorrect.

A graph of a sleep number

AI-generated content may be incorrect.

A graph of stress level

AI-generated content may be incorrect. A diagram of a distribution of bmi

AI-generated content may be incorrect. A graph with a line

AI-generated content may be incorrect.

A diagram of a clustering diagram

AI-generated content may be incorrect.

A group of colored dots

AI-generated content may be incorrect.

A diagram of colorful dots

AI-generated content may be incorrect.

**Exploratory Data Analysis**

The univariate distributions **(Figures 2–6)** were roughly symmetric and without values more than three standard deviations from the mean. The correlation matrix in **Figure 1** shows low associations between most pairs, except for stress and sleep, which present a moderate negative correlation (r = –0.42), and exercise and BMI, which reflect a moderate negative correlation (r = –0.35). However, all other pairings of indicators are quite low in magnitude, suggesting that each indicator is identifying a unique component of patient health.

**Determining the Optimal Number of Clusters**

Elbow plots, which are depicted in **Figure 7**, showed that within-cluster sum of squares (WCSS) decreases rapidly until k = 4 and then slowly but noticeably flattens. It means that after four clusters, the addition of more of them will produce less significant boosts. Ward’s linkage dendrogram in **Figure 8** also shows a large vertical gap prior to acquiring the main four branches, demonstrating that such several clusters represent the natural composition of the data and will not over fragment them. Thus, the combination of both visualizations supports using only 4 clusters to divide the patients’ wellness profiles.

**PCA Explained Variance**

The first 2 principal components account for most of the variability in the data set, 23.7% and 22.1% respectively for principal components (PC) 1 and 2 for a total of 45.8%, rendering a two‐dimensional representation much more manageable. Inspection of the component loadings shows that PC1 mainly opposes the frequency of healthy meals to stress, whereas PC2 opposes the duration of exercise to BMI. This simplified two‐component solution preserves major trends, reduces noise and multicollinearity, and provides a compact visual basis for further clustering.

**Cluster Profiles**

**Figure 9** displays the scatterplots of PC1 vs. PC2 with cluster assignments for (a) K-Means and (b) hierarchical clustering.

A screenshot of a graph

AI-generated content may be incorrect.We applied K-means with k = 4 to the original data and calculated the average values of these metrics per cluster:

* Cluster A (n = 60): High PA (M = 44 min), high healthy meals (M = 6/day), good sleep (M = 8 hrs.), low stress (M = 2), normal range BMI (M = 22).
* Cluster B (n = 46): Low physical activity (M = 25 min), few wholesome meals (M = 3/day), short sleep (M = 6 hrs.), high stress (M = 7), high BMI (M = 30).
* Cluster C (n = 48): Moderately active (M = 36 min), eating healthy (M = 5/day), sleeping enough (M = 7 hrs.), moderate stress (M = 5), moderately high BMI (M = 26).
* Cluster D (n = 46): High stress (M = 9), moderate exercise (M = 30 min), average sleep (M = 7 hrs.), low healthy meals consumed (M = 2/day), BMI = overweight (M = 28).

**Separation in Principal Component Space**

Figure 10 shows that each point is a patient represented in the 2d space of the first two components of a PCA, and the color indicates the cluster output of the K-Means algorithm. The blue cluster (0) is located on the left part of the plot (PC1 negative scores) while the orange cluster (1) is on the right, with positive high PC1 values. The green (2) and red (3) clusters are located at the high and low PC2 middle, respectively. Although they overlap in some degree, the general trend of clustering is apparent with the silhouette score of 0.355 suggesting well separated clusters at this reduced-dimensional space.

**Discussion:**

If we examine the silhouette scores, we observe that none of the methods achieved a perfect separation K-Means based on the full five variables achieved a score of 0.168, PCA-based K-Means achieved a score of 0.152, and hierarchical clustering achieved a score of 0.114. This modest score tells us that while the four groups are meaningfully different, there is still similarity at the edges. This is likely to mean that a small number of people do not easily fit into any given segment, and interventions would need to remain somewhat malleable.

Yet despite that overlap, the four clusters showed distinct patterns in health behaviors. Higher stress is associated with shorter sleep, for example, showing that mental well-being and sleep are closely related. We also found that people who tended to eat more of these healthy meals were also more often active, which indicates that upgrading in one area may help improve the other. By visualizing clusters in two dimensions-using PCA-we were able to present these findings in an interpretable form to non-technical stakeholders.

As the results are simulated, they would need to be revalidated on actual patient populations before introducing targeted programs. It is possible that adding covariates like age, sex, or medical history would have improved the quality of the segments. But longitudinal tracking following the same people over time would reveal whether people wander between clusters after interventions and help us measure true impact. In sum, this clustering strategy offers a solid foundation on which to build and tailor wellness efforts, but it needs to be confirmed and elaborated upon in practice.

**Recommendations for Improving the Wellness Program**

To make the wellness program more successful, due to begin with what you do elsewhere and offer individuals a variety of paths to support based on their existing human needs. For high-stressed, un-sleeping folk, weekly sleep tips and short guided relaxation sessions. For members who don’t work out much and don’t eat many healthy meals, they offer simple meal plans and beginner-friendly exercise videos. Get the most grounded group to spread success stories or guide others, to inspire a community. And finally, utilize a simple app or a printable chart to keep and track progress so that each family member can see the small wins and be encouraged to keep it going (Vuik, Mayer, & Darzi, 2016).

**How Clustering Helps Tailor Interventions for Patient Segments**

Clustering enables the patients with similar habits which type of support each group mostly needs. For instance, one company cluster may be poor at stress handling, and another may have no regular exercise. Once these cohorts are who they are, the health care team can send targeted messages to each — say, tips for reducing stress to one and exercise challenges to another — rather than one-size-fits-all communiqué. This targeted attitude is time and cost saving, and more personal, so patients are more likely to participate and stick to their schedule (Vuik et al., 2016).

**Conclusion:**

In this study, unsupervised learning techniques successfully characterized patient profiles of wellness for intervention studies. Based on the original feature set, K-Means clustering gave the best separation, while PCA-reduced clustering has similar performance with better visualization. There were four different categories of wellness that surfaced, and their program strategies were tailored -- from stress reduction workshop to peer-led fitness challenges. The segments identified in this study require validation in prospective patient cohorts, with inclusion of demographic and clinical outcomes, and longitudinal analysis of behavior change post-intervention.

**References**

Jolliffe, I. T. (2011). Principal Component Analysis (2nd ed.). Springer.

Kaufman, L., & Rousseeuw, P. J. (2009). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65.

Vuik, S. I., Mayer, E., & Darzi, A. (2016). A quantitative evidence base for population health: Applying utilization-based cluster analysis to segment a patient population. Population Health Metrics, 14, Article 44. https://doi.org/10.1186/s12963-016-0115-z