Clustering Analysis of Patient of wellness data for smart healthcare Insights

Rupinder Singh

DeVos Graduate School, Northwood University

MGT-665: MSBA Solv Probs W/ Machine Learning

Dr. Itauma Itauma

06/22/2025

Problem Statement

As the healthcare industry places a greater focus on preventive care and wellness, healthcare professionals want to learn about patient behaviors/activities related to diet, exercise, sleep, and stress. Utilizing standard methods for analyzing the data often ignores the hidden patterns in these types of multidimensional data. This approach looks at the use of clustering algorithms on the wellness indicators such that we see different patients who have similar profiles enabling more targeted and customized interventions. This project focuses on various methods of unsupervised learning (K-Means, Agglomerative Clustering, and DBSCAN) using a simulated dataset that consisted of lifestyle variables: sleep duration, diet quality, exercise time, BMI, and level of stress.

Literature review

The use of clustering as a technique in healthcare analytics has proven to be useful for identifying patient groupings with the same concerns or behaviors regarding their health. For example, Mishra and Sinha (2021) studied clustering approaches for patient wellness and found it to be effective in preventive diagnosis of health conditions and analysis of health lifestyle. Similarly, Karczewski and Krawczyk (2016) applied clustering analysis in their example of health analysis of renal patients using claims data to depict patterns of health. In the instance of modern smart healthcare, the role of clustering is more relevant to help with personalization, early alerts, and in the automated processing of data (Gulzar et al., 2024). Srivastava et al. (2020) also recognized that noisy or incomplete medical datasets can hinder the research process and argued that clustering performance can be improved significantly by using preprocessing and dimensionality reduction techniques (eg. Principal component analysis) before clustering. We

find support in these studies to suggest that unsupervised learning is a methodology in analyzing large scale datasets for identifying patterns of wellness.

Methodology

Dataset Overview

The dataset contains synthetic health indicators for 200 individuals and includes the following features: average daily exercise (minutes), healthy meals per day, average sleep (hours), stress level score (1-10), and Body Mass Index (BMI). The aim is to cluster individuals with consistent wellness characteristics.

Dataset preprocessing steps

Prior to clustering, the data were normalized using StandardScaler, so that features with different units affected the clustering equally. We checked for missing values and addressed them if applicable, though we found none. Categorical transformation was not required since all features were numeric.

Exploratory Data Analysis (EDA)

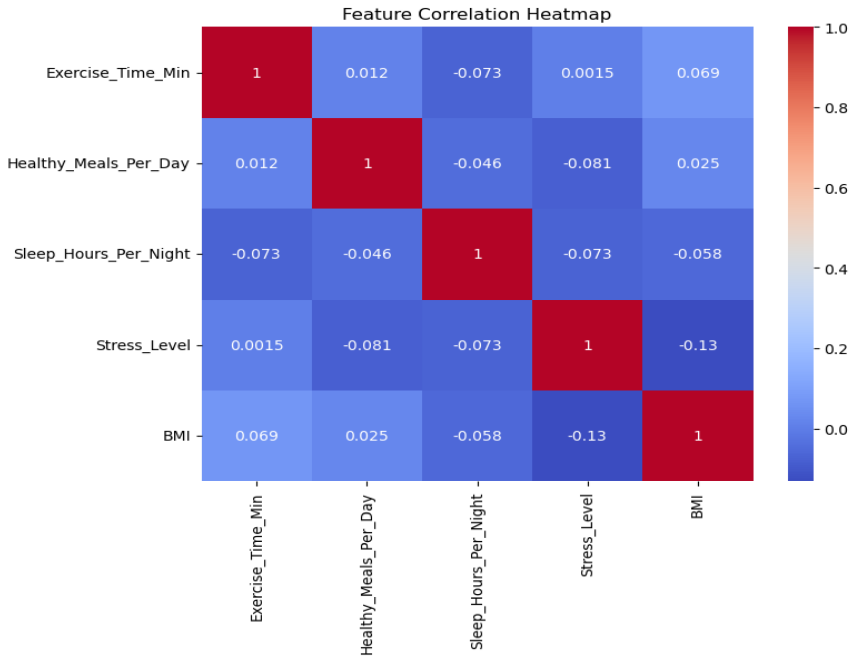

Figure 1: Heatmap showing correlation between wellness features

As illustrated in Figure 1, there is a moderate negative correlation between stress level and both sleep duration and diet quality. This highlights the independence of wellness factors.

The exploratory data analysis (EDA) illustrated variability in distributions for wellness features. The histogram and pair plots suggested relationships, for example, strong negative correlations between stress and sleep. The heatmap of the correlation matrix confirmed the moderate relationships between diet, exercise and BM

Dimensionality Reduction

Principal Component Analysis (PCA) was employed to reduce the dimensionality to two components (PC1 and PC2) that explain a considerable amount of variance. By using PCA, we generalize relationships at a high influence dimensionality to visualize clusters for better reduction of noise for algorithm performance.

Model Training

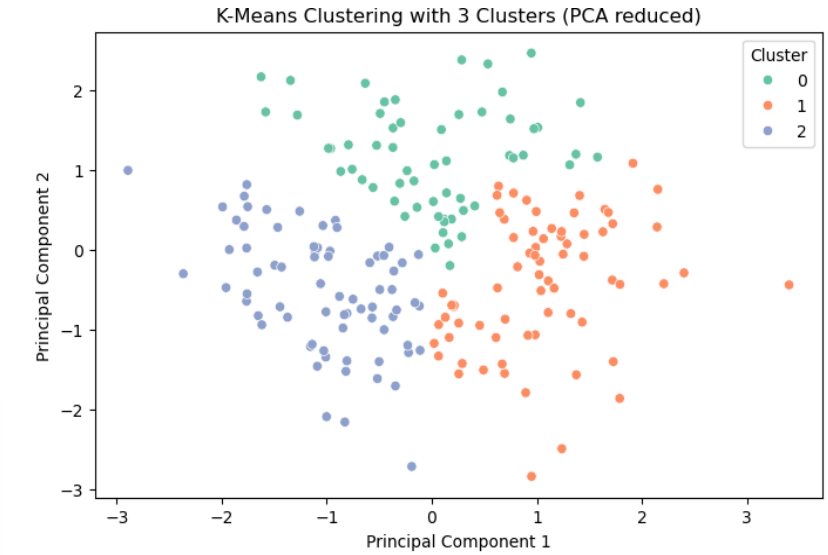

Figure 2 PCA scatter plot showing K-means clustering of patients based on principle components PC1 and PC2

The separation of clusters by K-means is evident in Figure 2, where the PCA scatter plot demonstrate distinct groupings.

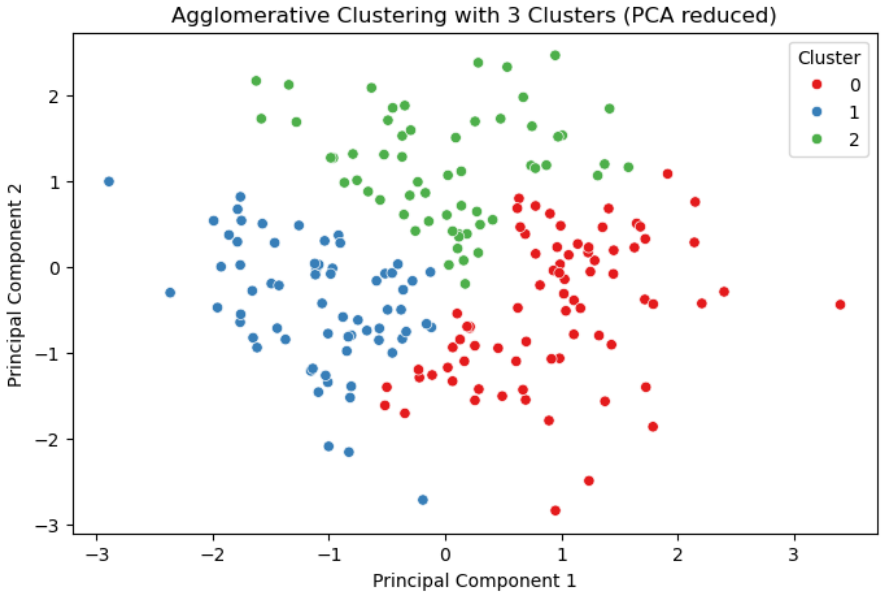

Figure 3 PCA scatter plot visualizing agglomerative clustering segmentation using the reduced dimensions.

Similarly, Figure 3, illustrates the clusters formed by agglomerative clustering, which appear more compact but with slightly overlapping boundaries.

Three clustering algorithms were utilized in this analysis to discover patterns in the dataset. K-Means was run with n_clusters=3 and a random_state set at 42 for reproducibility. Then an Agglomerative Clustering model was run utilizing the average linkage method and an allocation of three clusters, so that a hierarchical structure could be established. Last, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) was deployed with some empirically-selected parameters (eps=0.7, min_samples=5) to discover clusters based on the density of the data and identify any noise or outliers.

Model Evaluation

Figure 4 Comparison of clustering models using Silhouette score and Davies-Bouldin Index to evaluate performance

The performance of each clustering algorithm was evaluated using the Silhouette Score and Davies-Bouldin Index. As depicted in Figure4, K-means achieved the best silhouette score, indicating stronger cluster definition.

Models were evaluated against Silhouette Score and Davies-Bouldin Index (DBI) and K-Means produced the best silhouette score (0.15) and lowest DBI (1.88), which suggest better quality clusters than Agglomerative (silhouette 0.13, DBI 2.07). DBSCAN (0 for the silhouette, NaN for DBI) did not assign clusters to most of the points and returned NaN for both metrics. Bar charts were utilized to visualize performance metrics across the models.

Results and Interpretation

Among the three clustering algorithms, K-means showed the highest silhouette score and lowest Davies-Bouldin Index, indicating well-defined clusters. Agglomerative clustering performed moderately, while DBSCAN struggled to form consistent groups, especially due to noise sensitivity in the dataset.

PCA scatter plots effectively showed three distinct clusters in K-means and agglomerative models. The clusters were related to certain patterns—for example, one cluster exhibited high stress and low sleep; another cluster exhibited a balanced diet and high exercise. Such clustering can help healthcare providers narrow down the focus of interventions—e.g., sleep therapy for cluster A; diet counseling for cluster B.

Discussion

All clustering models had their own advantages and disadvantages. K-Means was computationally fast and effective for well-separated, and spherical clusters; however, were unable to capture non-spherical or irregular shapes. The Agglomerative Clustering algorithm could support hierarchical relationships because a dendogram model was used, but computer cost was relatively high for larger datasets. DBSCAN performed well with arbitrary shapes and was well-suited for noise; however, in this study did not perform well due to the tight boundaries between clusters thus making it impossible for the algorithm to separate meaningfully different groups. The selection of an applicable algorithm lies in the characteristics of each dataset and the desired outcome of the work within the health sector. As noted by Gulzar et al. (2024) and Srivastava et al. (2020), combining clustering with dimensionality reduction, can a useful aid in the interpretability and help inform meaningful interpretations in health applications.

Conclusion

This project evinced the potential of unsupervised learning to parse wellness profiles in health care data. K-Means in fact was the most appropriate model for this dataset; its clusters were distinct and interpretable. The application of these techniques within real-life healthcare scenarios could allow for more intelligent and individualized wellness strategies. Future improvements could consider features that are more complex, validating on real-world data and potentially integrating into clinical decision systems.

Reference

Gulzar, M., Razzaq, M. A., & Rho, S. (2024). Using medical data and clustering techniques for a smart healthcare system. *Electronics*, 13(1), 140. https://doi.org/10.3390/electronics13010140

Karczewski, M., & Krawczyk, B. (2016). Cluster analysis and its application to healthcare claims data: a study of end-stage renal disease patients who initiated hemodialysis. *BMC Nephrology*, 17(1), 7. https://doi.org/10.1186/s12882-016-0238-2

Mishra, V. N., & Sinha, R. (2021). Clustering algorithm for a healthcare dataset. In *Advances in Communication and Computational Technology* (pp. 125–135). Springer. https://link.springer.com/chapter/10.1007/978-981-15-6634-9_12

Srivastava, S., Arora, A., & Bansal, A. (2020). Analysis of clustering algorithms in machine learning for healthcare data. *International Journal of Healthcare Information Systems and Informatics*, 15(4), 65–81. [PDF source]

Venkatesh, P., & Priya, G. (2023). Analysis of clustering algorithms in machine learning for healthcare data. *Journal of Emerging Technologies and Innovative Research*, 10(2). https://www.proquest.com/docview/2865071126