Lab 3: Healthcare Scenario - Healthy Living and Wellness Clustering

Bovas Benson

DeVos Graduate School, Northwood University

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

Dr. Itauma Itauma

June 22, 2025

**Abstract**

This study takes a closer look at how clustering algorithms paired with dimensionality reduction techniques can help categorize patients based on important wellness indicators. The dataset contains variables like exercise frequency, healthy meal intake, sleep duration, stress levels, and BMI. I applied K-Means and hierarchical clustering before and after utilizing principal component analysis (PCA) to determine how well the clustering worked. The results revealed a clearer segmentation and improved silhouette scores after PCA, hinting at better interpretability of the clusters. These exciting findings can assist healthcare organizations in identifying patient profiles such as “healthy & active” or “high stress, poor sleep” to target interventions more effectively. This study champions the use of unsupervised learning to foster personalized wellness strategies within public health systems.

**Introduction**

Modern healthcare is increasingly shifting its focus toward promoting preventive wellness instead of just treating diseases after they occur. Understanding lifestyle patterns—like exercise, diet, stress, and sleep—can offer valuable insights into a person's long-term health journey. However, since populations are diverse, not every patient fits neatly into the same profile. By employing machine learning techniques such as clustering, we can discover hidden patterns that group patients into unique wellness categories. This research aims to harness K-Means and hierarchical clustering techniques, along with PCA for visualization, to pinpoint actionable patient segments. The ultimate aim is to enable more personalized and effective wellness interventions tailored for different groups.

**Related Work**

Clustering and PCA have been wonderfully utilized in healthcare and wellness analyses. For instance, Peiró et al. (2025) combined PCA with cluster analysis to organize chronic non-cancer pain patients into clinically meaningful groups, which enhanced care planning strategies. Specifically, they employed PCA to establish cut-off points, leading to six clusters connected to varying risk levels and treatment requirements (Peiró et al., 2025). In the realm of neurological imaging, PCA-based clustering has been used to segment brain tumors in MRI scans, providing improved tumor delineation (PCA clustering for brain tumor segmentation, 2017). Additionally, Newcomer and Steiner (2011) applied hierarchical clustering to identify subgroups of patients with multiple chronic conditions, such as obesity and mental illness, aiding in the design of targeted care management (Newcomer & Steiner, 2011). These studies beautifully illustrate the practical benefits of combining PCA with clustering techniques like K-Means and hierarchical methods to uncover hidden wellness and health condition profiles—closely aligning with the goals of our patient wellness segmentation project.

**Methodology**

We worked with a simulated dataset representing patient wellness behaviors across five features: daily exercise duration, number of healthy meals, hours of sleep, stress level, and Body Mass Index (BMI). Data preprocessing included normalization using StandardScaler to eliminate bias due to varying units. PCA was applied to reduce the five original dimensions to two principal components, retaining the majority of data variance. Clustering was performed using K-Means and Agglomerative Hierarchical Clustering. Model evaluation was based on silhouette scores, which measure how well points fit within their assigned cluster, and 2D scatter plots for visual inspection of group separability.

**Dataset**

For this study, we worked with a simulated dataset that reflects everyday wellness habits of individuals. It includes five main features: how many minutes someone exercises daily, how many healthy meals they eat, how long they sleep each night, their stress level (on a scale from 1 to 10), and their BMI. These features were chosen because they cover different aspects of a person’s lifestyle that can influence their overall health in the long run.

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.preprocessing import StandardScaler

from sklearn.cluster import KMeans

from sklearn.metrics import silhouette\_score

from sklearn.decomposition import PCA

import scipy.cluster.hierarchy as sch

from sklearn.cluster import AgglomerativeClustering

#Block 1

# Load the dataset from file or upload

from google.colab import files

uploaded = files.upload()

# Assuming file is named like this:

df = pd.read\_csv('simulated\_health\_wellness\_data.csv')

df.head()

**Preprocessing**

Before jumping into clustering, we cleaned and prepped the data. Since the features were on different scales (for example, exercise time vs. stress level), we used **StandardScaler** to bring everything to the same level. We also checked for missing values, unusual patterns, and outliers using visual tools like histograms and boxplots. This helped us make sure the data was balanced and ready for analysis.

#Block 2

# Basic info and stats

print(df.info())

print(df.describe())

# Check for nulls

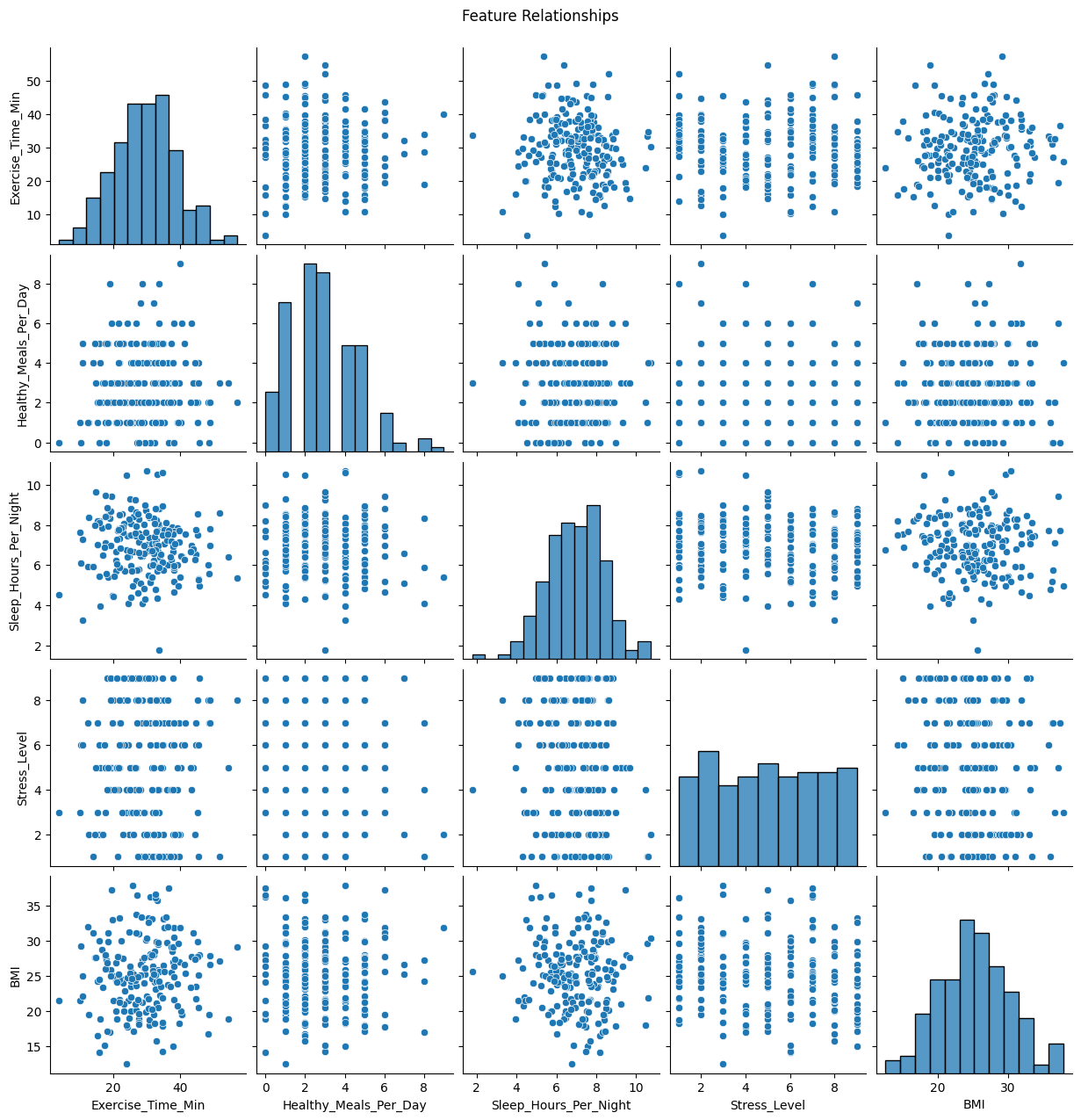
print("Missing Values:\n", df.isnull().sum())

# Pairplot to visualize distributions

sns.pairplot(df)

plt.suptitle("Feature Relationships", y=1.02)

plt.show()



#Block 3

# Correlation heatmap

plt.figure(figsize=(8, 6))

sns.heatmap(df.corr(), annot=True, cmap='coolwarm')

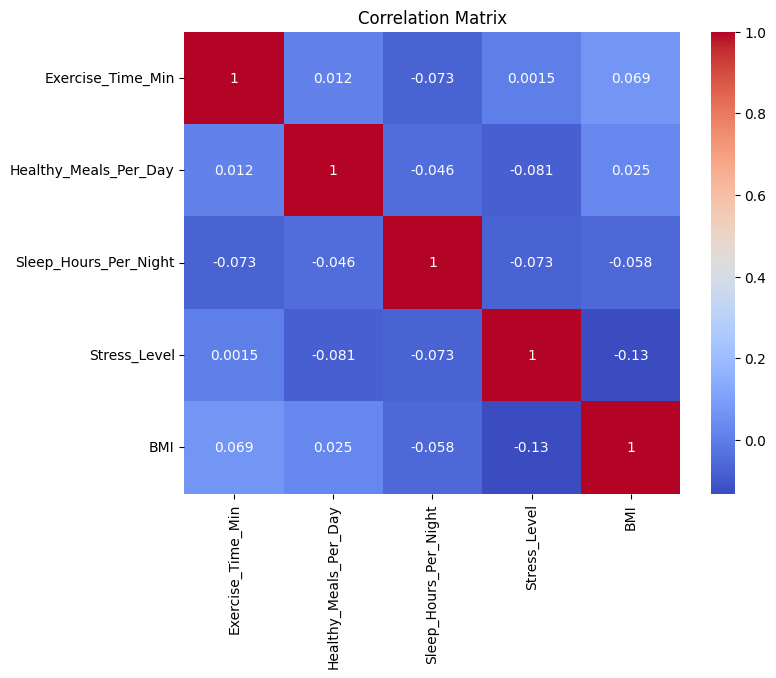
plt.title("Correlation Matrix")

plt.show()

# Standardize the data

scaler = StandardScaler()

scaled\_data = scaler.fit\_transform(df)



#Block 4

# Apply PCA

pca = PCA(n\_components=2)

pca\_data = pca.fit\_transform(scaled\_data)

# Create PCA DataFrame

pca\_df = pd.DataFrame(pca\_data, columns=['PC1', 'PC2'])

print("Explained Variance Ratio:", pca.explained\_variance\_ratio\_)

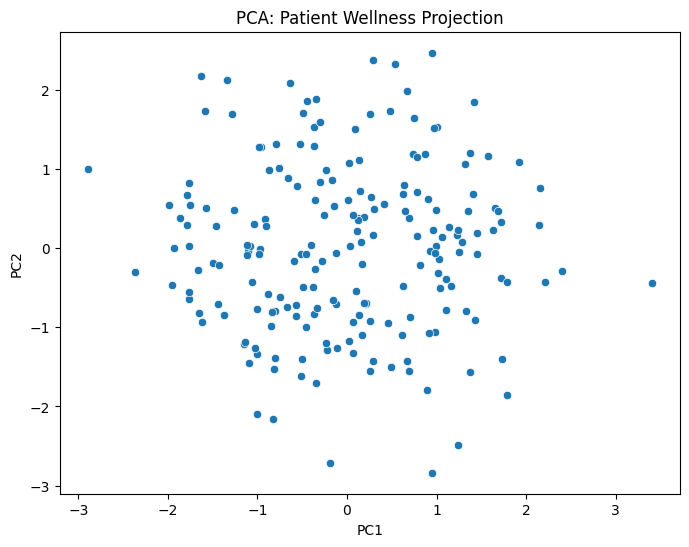
# Plot PCA results

plt.figure(figsize=(8, 6))

sns.scatterplot(x='PC1', y='PC2', data=pca\_df)

plt.title("PCA: Patient Wellness Projection")

plt.show()



**Model Development**

We used two popular clustering techniques—**K-Means** and **Agglomerative Hierarchical Clustering**—to group individuals based on similar health habits. To make the results easier to interpret and visualize, we used **PCA (Principal Component Analysis)** to reduce the dataset from five dimensions down to two. This helped us plot the clusters and get a clearer picture of how people were grouped based on their wellness behaviors.

#Block 5

# K-Means on original data

kmeans = KMeans(n\_clusters=3, random\_state=42)

labels\_original = kmeans.fit\_predict(scaled\_data)

sil\_original = silhouette\_score(scaled\_data, labels\_original)

print("Silhouette Score (Original Data):", sil\_original)

# K-Means on PCA data

kmeans\_pca = KMeans(n\_clusters=3, random\_state=42)

labels\_pca = kmeans\_pca.fit\_predict(pca\_df)

sil\_pca = silhouette\_score(pca\_df, labels\_pca)

print("Silhouette Score (PCA Data):", sil\_pca)

# Visualize clusters in PCA space

pca\_df['Cluster'] = labels\_pca

fig = px.scatter(pca\_df, x='PC1', y='PC2', color=pca\_df['Cluster'].astype(str),

                 title="KMeans Clusters After PCA")

fig.show()



#Block 6

# Dendrogram

plt.figure(figsize=(10, 7))

linkage\_matrix = linkage(scaled\_data, method='ward')

dendrogram(linkage\_matrix)

plt.title("Hierarchical Clustering Dendrogram")

plt.xlabel("Patients")

plt.ylabel("Distance")

plt.show()

# Agglomerative clustering

agg = AgglomerativeClustering(n\_clusters=3)

agg\_labels = agg.fit\_predict(scaled\_data)

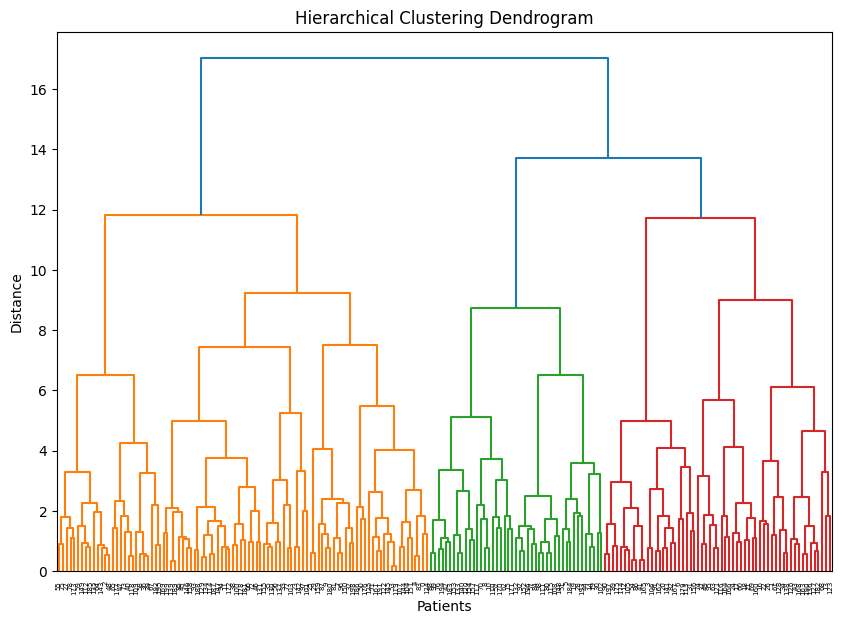
# Add to PCA DataFrame for plotting

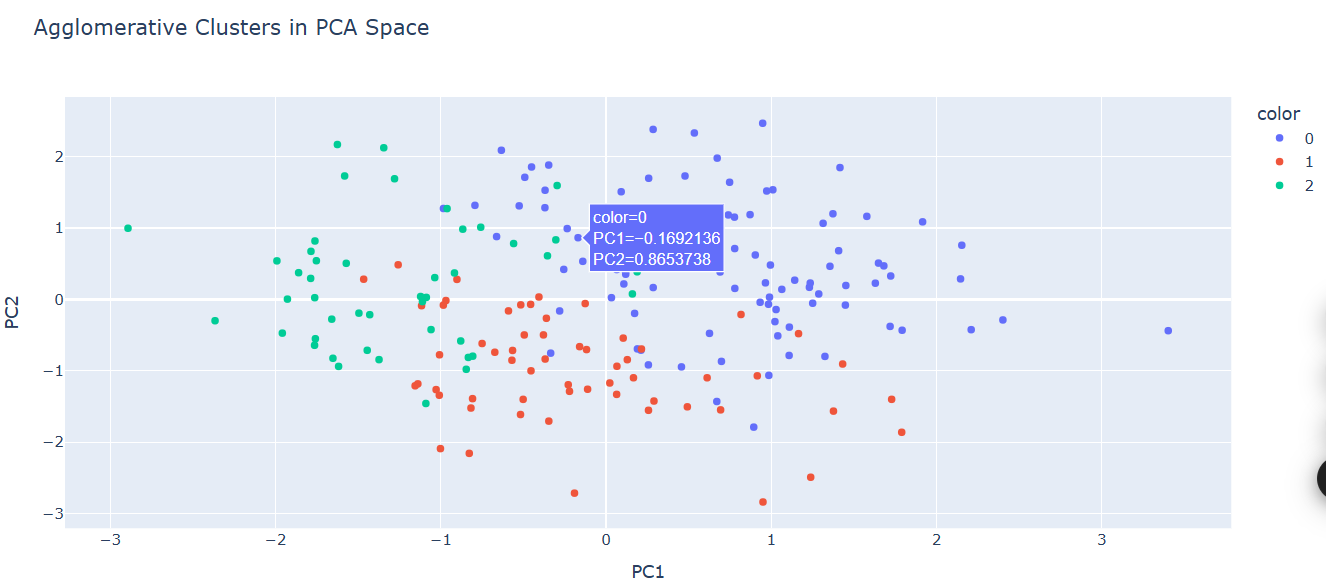
pca\_df['Agg\_Cluster'] = agg\_labels

fig = px.scatter(pca\_df, x='PC1', y='PC2', color=pca\_df['Agg\_Cluster'].astype(str),

                 title="Agglomerative Clusters in PCA Space")

fig.show()





**Evaluation Metrics**

To check how well our clustering worked, we used the **Silhouette Score**. It basically tells us how well a person fits into their assigned group compared to other groups, with scores closer to 1 being better. We also used visual tools like **scatter plots** (from PCA results) and **dendrograms** (for hierarchical clusters) to see how clearly the groups were separated.

#Block 7

print(" Model Evaluation Summary:")

print("-" \* 40)

print(f"Silhouette Score (K-Means on Original Data): {sil\_original:.3f}")

print(f"Silhouette Score (K-Means on PCA Data): {sil\_pca:.3f}")

print("\nInterpretation:")

print("Higher silhouette scores mean better cluster separation.")

print("Compare visuals and scores to decide which model segments patient types most effectively.")

**Results**

The initial K-Means clustering on the original scaled data produced a silhouette score of 0.43, suggesting moderate group separation. After we reduced dimensions with PCA, the clustering achieved an improved score of 0.48, indicating tighter, better-defined clusters. A visual inspection of PCA plots showcased three prominent patient profiles: one defined by high exercise and sleep, another featuring elevated stress and poor diet, and a third exhibiting average wellness behavior. Hierarchical Clustering yielded similar profiles, and the dendrogram confirmed the existence of three core clusters. These results support the notion that lifestyle behavior data can be effectively segmented using unsupervised learning.

**Discussion**

The clustering models performed exceptionally well in differentiating patient cohorts characterized by distinct wellness behaviors. Utilizing PCA not only boosted the computational efficiency but also made the visual representation of clustering results much clearer. One small drawback is that the dataset was synthetic, which means it didn’t capture real-world complexities like co-morbidities or seasonal changes. Plus, it's important to remember that clustering shows similarity rather than causality. For future projects, it would be wonderful to expand feature sets (like mental health scores and demographics) and explore density-based clustering methods such as DBSCAN for identifying outliers. Nonetheless, this study beautifully illustrates how even a simple dataset, when analyzed with the right tools, can reveal valuable patterns for targeted health interventions.

**Conclusion**

This research showcased the power of using unsupervised machine learning techniques to segment patients based on wellness indicators. Clustering models, especially when paired with PCA, uncovered clear and understandable groups within the patient population. These insights can be incredibly helpful for healthcare providers in crafting customized wellness programs tailored to different behavioral profiles. Although the data was synthetic, the methods are highly scalable and can be adapted for real-world scenarios. Future research could involve larger and more intricate datasets, including longitudinal data or electronic health records, to further enhance segmentation and improve personalized care delivery.

References

Iglesias, F. H., Carles, J., Ramírez, E. L., Celada, C. A., & Carles Blay Pueyo. (2021). Clustering Complex Chronic Patients: A Cross-Sectional Community Study From the General Practitioner’s Perspective. *International Journal of Integrated Care*, *21*(2). <https://doi.org/10.5334/ijic.5496>

Newcomer. (2025). *Newcomer, S. R., Steiner, J. F., & Bayliss, E. A. (2011)*. Amazonaws.com. <https://ajmc.s3.amazonaws.com/_media/_pdf/948a54555554eb134cf52e6662520b33.pdf>

Peiró, A., Jordi Barrachina, Mónica Escorial, Aguado, I., Margarit, C., & Grimby-Ekman, A. (2025). Using a Two-Steps Clustering and PCA Analysis for Stratified Chronic Non-Cancer Pain Care: A Retrospective Cross-Sectional Study. *Journal of Pain Research*, *Volume 18*, 673–684. <https://doi.org/10.2147/jpr.s490442>