Lab 3: Healthcare Scenario - Healthy Living and Wellness Clustering

Priyal Rawat

DeVos Graduate School, Northwood University

MGT 665: Solving Bus Problems W/ Machine Learning

Dr. Itauma Itauma

06/22/2025

Abstract

This study focuses on the application of unsupervised machine learning algorithms to identify wellness profiles from a simulated medical dataset. The dataset consisted of 200 patients details including their five key lifestyle variables such as exercise time, healthy meals per day, hours of sleep, stress level and body mass index. Unsupervised machine learning techniques like K-Means, Agglomerative Hierarchical Clustering, Gaussian Mixture Models (GMM), and DBSCAN were performed for clustering and identifying patterns. Dimensionality reduction was done using Principal Component Analysis (PCA), which provided more insight into the data. Evaluation metrics which were used as Silhouette Score, Within-cluster Sum of Squares (WCSS), Davies Bouldin Index and Calinski Harabasz Score to determine a number of well-formed clusters. K-Means was identified as the best technique, achieving the highest silhouette score of 0.398 after PCA transformation, and showing well-formed, distinct clusters. The findings of this study provide support for the use of clustering models to enable the segmentation of patient populations in support of wellness intervention strategies. The study illustrates how unsupervised learning may be leveraged to accelerate data-driven approaches to preventive healthcare.

Introduction

Health is Wealth! It is the the most valuable aspect of human life. It plays a significant and critical role not only in individual well being, but it also influences the productivity and resillience of entire communities. In the last couple of years, the focus on health and preventive health has become more prominent as awareness of chronic diseases, mental health challenges, and unhealthy behaviours has grown. Health care professionals need to not only assess the original state of health of the patient, but they need to understand the complex picture of health,

to develop meaningful wellness programs and prescribe appropriate health interventions. The amount of health data is growing in both size and complexity. Therefore, the need for intelligent methods to analyze and interpret actual health data meaningfully is critical now and going to be an essential element in the future.

Unsupervised machine learning has become a revolutionary methodology for analysis in health-related systems and healthcare. Unsupervised techniques explore and cluster observations regardless of labels or categories, highlighting hidden structures in complex datasets (Alanazi, 2022; Jayatilake & Ganegoda, 2021). Clustering methods, such as K-Means and Hierarchical Clustering, have been used to explore lifestyle patterns while patients are grouped together according to similar behaviors. Other methods of dimensionality reduction, such as Principal Component Analysis (PCA) combine measures to help simplify data in a higher dimensionality and for ease of interpretation. These methods are necessary for capturing the complexity of health and wellness.

This study focuses on supporting a healthcare organization in understanding a simulated dataset that includes amount of daily exercise, diet quality, sleep duration, stress level scores, and Body Mass Index (BMI). Through clustering, patients are divided into wellness categories for targeted initiatives. Dimensional reduction technique like PCA has been used to how it impacts clustering, and the clarity of the patient groupings. The insights from this project could support a more individualized or personalized wellness approaches rather than delivering general wellness recommendations (Trezzza et al., 2024; Allenbrand, 2024).

This study is significant as it contributes to healthcare by an approach of using population level data to develop personalized, tailored treatment strategy based on an individual's unique pattern of data. By identifying wellness clusters from a single population group, a healthcare

system can more effectively allocate resources knowing the risk levels. Also, a healthcare system can deliver programs that are specific to the lifestyle components of the individual. Prior research has also indicated that the methods of this current research have strong outcomes in the implementation of drugs, fraud prevention and prediction of diseases (Lu & Uddin, 2024; massi et al, 2020). Integrating clustering and PCA is an important advancement towards a more intelligent, smarter, and flexible responsive healthcare system.

Literature Review

The implementation of machine learning in healthcare has increased rapidly, as healthcare organizations have large amount of patient data, both structured and unstructured data. This has changed healthcare industries from being reactive being proactive by identifying data driven opportunities. Also, a wide body of research has been done, looking at the importance of machine learning in clinical decision-making, disease diagnosis, and individualized treatment in both supervised and unsupervised studies, (Jayatilake & Ganegoda, 2021). Although many publications focused on disease prediction using labeled datasets, recent literature has proposed that unsupervised approaches can be helpful in finding patterns in unlabeled health data to support healthy behavior initiatives.

Unsupervised learning, including clustering methods, is becoming a viable option for patient segmentation. Trezza et al. (2024) suggest that unsupervised learning is significant for precision medicine because it can reveal hidden subgroups of patients without the requirement of labels already established. The authors state that unsupervised learning is valuable for personalizing care pathways, mainly in lifestyle and behavioral health pathways. Allenbrand (2024) utilized both the machine learning approaches, supervised and unsupervised to analyze data from health and wellness around pharma and medications. The diversity in results indicates

that unsupervised methods can be extremely significant and show incredible results across various domains. These studies show the ways clustering algorithms, such as K-Means or Hierarchical Clustering, may be applied in segmentation of the many behaviors and attitudes in health.

Pu and Uddin (2024) contributes by doing a comparative study of unsupervised methods using multiple databases in health care provides another contribution. Their study illustrates many clustering methods, differences identified based on the data context and dimensionality, and the importance of feature reduction intervention like Principal Component Analysis (PCA), which assists with both computational reduction and interpretation of clusters, particularly in domains of wellness data containing many inter-related variables such as sleep, stress, nutrition. However, Massi et al. (2020) used a two-step clustering process to identify outliers in administrative databases for the purposes of fraud analysis. Their example shows that clustering provides a value beyond just clinical diagnostic, such as clustering in the health and wellness domain.

Alanazi (2022) provides a comprehensive overview of machine learning in health care addressing health issues with a point to future work indicating the number of wellness studies either in association with real-world data or simulated data, as most of the literature in this area is related to disease prediction and improving forms of treatment approaches. However, some literature focusses explicitly on clustering patients, relating specifically to wellness data, and that could inform potential preventive wellness programming, but on a large dataset. While this study is inspired by other literature, it addresses the important gap by applying unsupervised learning techniques to a narrow data set made up of core wellness variables which are, exercise time, eating behavior, sleep, stress levels, and BMI. Instead of analyzing a broad-spectrum patient

attribute study, this analysis focuses on everyday conduct factors that are changeable, directly related to general wellness.

To sum up, while prior research supports the utility of unsupervised learning in healthcare, there is still opportunity to investigate the applicability to lifestyle segmentation in wellness-type datasets. This study expands on the work of Trezza et al. (2024) and Lu and Uddin (2024) but does so with a more focused dataset on actionable health behaviors. This offers a new insight into how health organizations can utilize clustering and dimensionality reduction techniques to tailor wellness interventions and promote healthier populations.

Methodology

Dataset Overview

The dataset for this study has 200 records of wellness profiles, based on five central lifestyle measures. These measures are denoted as daily exercise time in minutes, the number of healthy meals consumed per day, hours of sleep each night, and participants self-reported their overall stress level on a scale from 0 to 10 and body mass index (BMI). This concentrated dataset provides key modifiable factors related to health, as it incorporates central ways in which patients can take control over their health outcomes. This dataset gives an insight into patients day to day life behaviour which is quite different from clinical dataset that focuses majorly on disease, medical record or disease specific details. The dataset focuses on regular health behaviours, thus can be effectively analysed using unsupervised learning methods to extract additional, useful insights or to create patient profile segments using clustering approaches. It can be further refined and explored successfully at each level of clustering and dimensionality reduction, to assess and describe wellness pathways and help design specific health interventions.

Handling Missing Values

The first stage was examining the dataset for null values. A column-wise examination confirmed that there were no null entries across all variables, allowing the dataset to be used in its entirety without the use of imputation. Then data types and exploratory analysis was done to assess for inconsistencies or outlier values.

Table 1

Missing Values

Variables	Missing Values
Exercise_Time_Min	0
Healthy_Meals_Per_Day	0
Sleep_Hours_Per_Night	0
Stress_Level	0
BMI	0

Note: Table 1 displays the number of null values discovered for each variable of the dataset. As illustrated, no missing data were discovered for any of the five health and wellness indicators, which allowed the complete dataset (N = 200) to be used in the analysis without proceeding with imputation.

Exploratory Data Analysis

To gain additional insights into the dataset, an exploratory data analysis (EDA) was performed on features such as exercise time, healthy meals consumed each day, sleep hours, stress level, and BMI. All features and data types were analyzed separately and all 5 features began with a histogram for each features to show the distribution of the feature. The distributions for exercise time, sleep hours, and BMI were approximately normal. The distribution for healthy meals consumed per day was positively skewed and therefore suggests that respondents select

more healthy meals than average, the distribution for stress level appeared to have no clear distribution pattern. After obtaining histograms, a correlation heatmap was created to view correlations or relationships among the variables. Overall, the features were mostly weakly correlated with each other, which may indicate that these features may contribute in different ways to the latent variable of patient segmentation. The most negative correlation noted was between stress level and BMI ($r = –0.13$).

A scatter-KDE pairplot was created to examine the interactions and separability of variables. This offered distributions of the singular variables, along with pairwise relationships, but also provided some evidence of clustering in discrete combinations such as sleep and stress or exercise and BMI. A swarm plot was used next to show the actual distribution of each individual data point, revealing repeat measurements of discrete variables such as healthy meals and stress level, while exposing outliers of BMI and other continuous features. A violin plot, in contrast, provided a smoothed, statistical representation of the distribution of each feature, in which a kernel density estimation was used to provide patterns of spread, skewness, and central tendency with boxplots imbedded within the plot. Together, these two graphs offered complementary visualizations and built a strong analytical foundation for clustering, revealing possible locations of natural behaviors in the patient population.

Figure 1

Distribution of Features

Note: Figure 1 shows the distribution of all five wellness indicators across the 200 patients in the dataset. Exercise time, sleep time, and BMI are roughly normally distributed, whereas healthy meals per day is right-skewed, reflecting that few people eat a high number of healthy meals. Stress level seems to be uniformly distributed, reflecting that the sample population experiences stress differently.

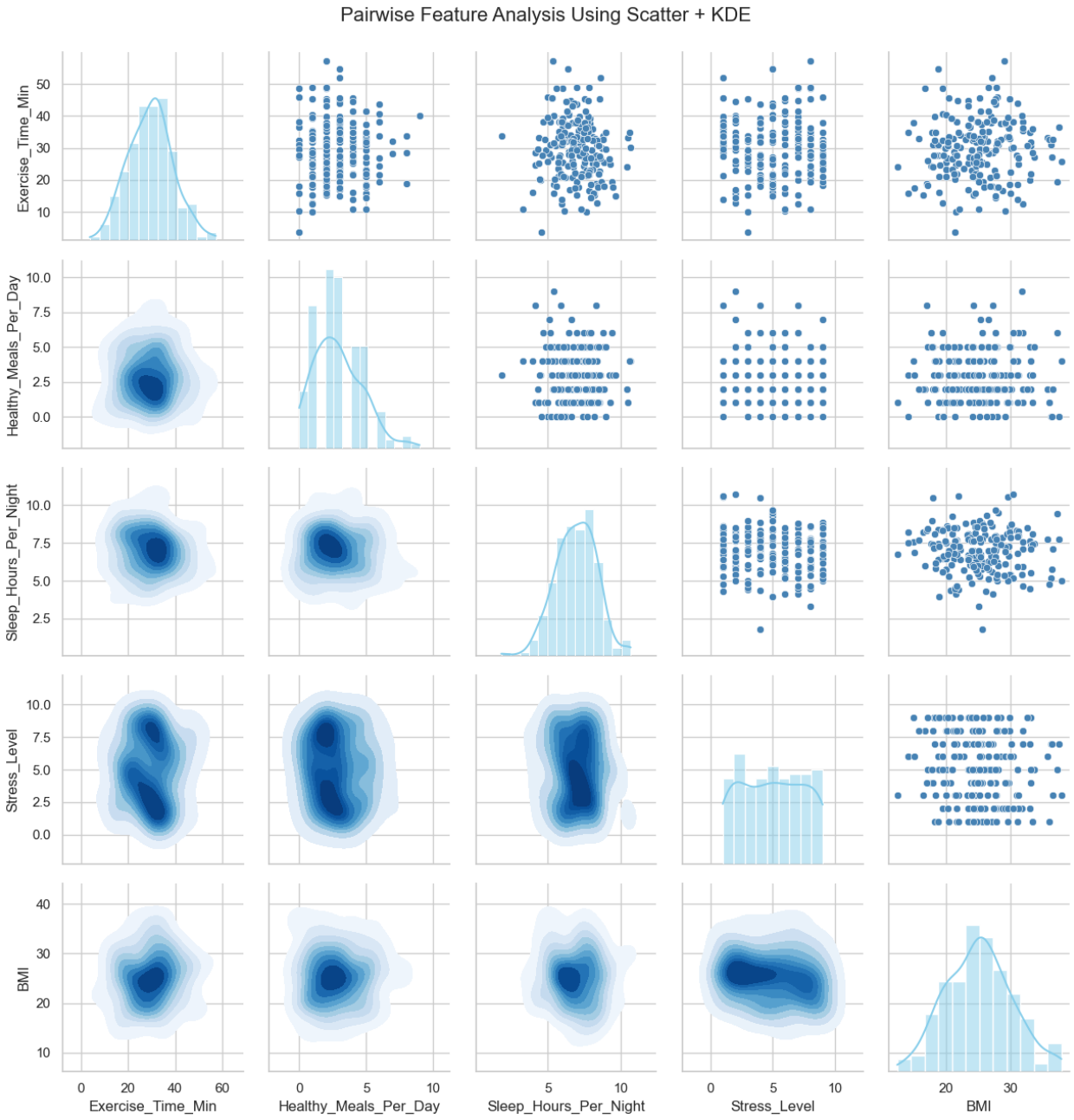
Figure 2

Correlation Heatmap

Note: Figure 2 correlation heat map displays Pearson correlation coefficients between the five wellness variables. There are mostly weak correlations, with the strongest negative relationships being between stress level and BMI (r = -.13). Although exercise time, sleep time, and healthy meals per day have low correlations with each other, this suggests that one or more of these variables may be independent of one another or at least not related through behavior.

Figure 3

Pairwise Feature Analysis Using Scatter and KDE Plots

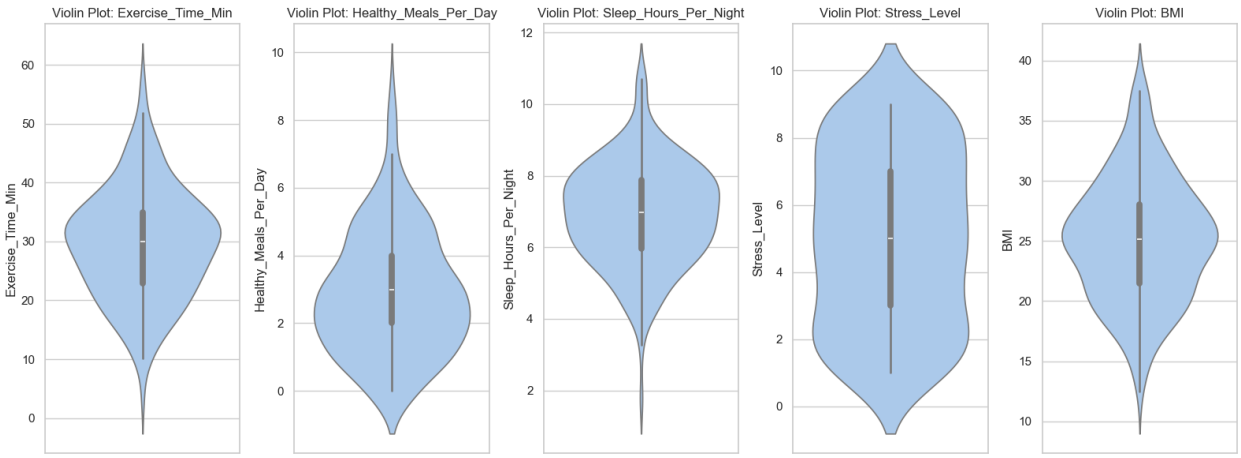


Note: Figure 3 shows pairwise relationships between all health features with scatterplots kernel density estimates, and histograms. Most combinations of features don't have linear relationships,

but BMI and exercise time have some density of pattern. The smooth contours indicate the density of observations, which provide evidence for locating potential clustering structures.

Figure 4

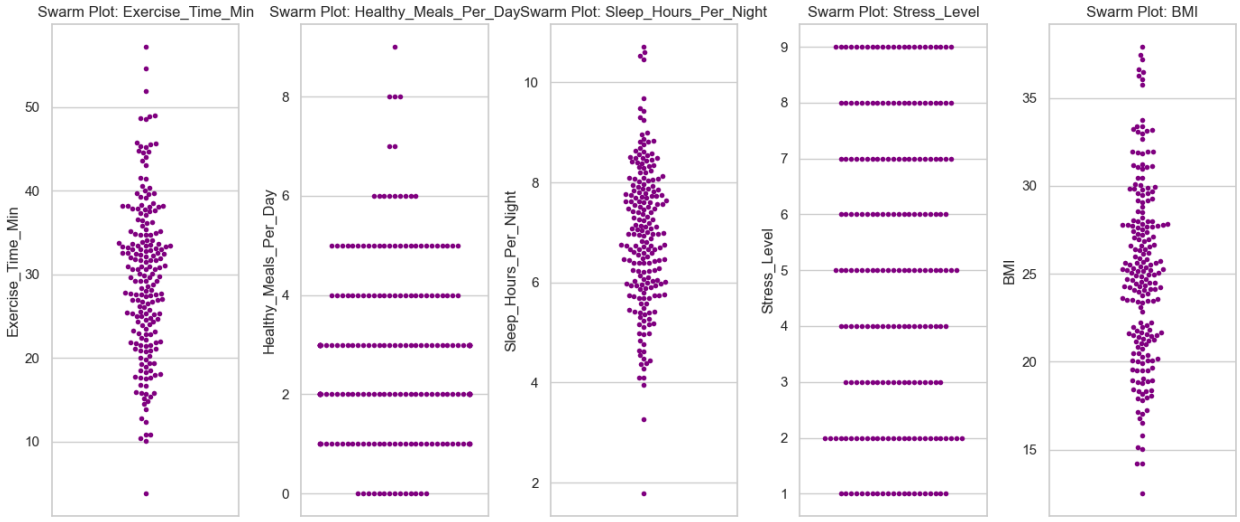
Violin Plot of features



Note: Figure 4 Violin plots show the distribution and density of each feature related to wellness using a boxplot that also incorporates a kernel density estimate. The inner white dot represents the median, while the overall shape portrays the spread of the data and skewness. The distributions related to exercise time and BMI tended to have moderate symmetry, while other features like stress level spread out more uniformly in one direction.

Figure 5

Swarm Plot



Note: The plot clearly reveals columned values of individual data points for each wellness feature, which highlights the points and the density and repetitions of certain values. The plots containing discrete variables like number of healthy meals per day and stress level also show stacking patterns of values due to how many times the values were repeated across the sample. The plots containing continuous variables like BMI and sleep duration contrast these stacking behaviors of discrete variables demonstrating its natural spread and where there is natural clustering around the value. The graphs also give a chance to examine outliers directly and what the distributions of individual records look like over the sampled measures.

Standardization

Both K-Means clustering, DBSCAN clustering, and Principal Component Analysis (PCA) performs better if there are not any major differences in scale of the features. Thus, every variable was standardized Z-Score normalization using the StandardScaler from scikit-learn. The features, on average, were then rescaled and have a standard deviation of 1 and a mean of 0,

meaning that every variable will contribute, equally, as part of the distance calculations used in clustering algorithms.

Table 2

Dataset after Standardization (Top 5 Rows)

	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	-1.53711	-0.27101
3	1.683908	-1.035394	1.149993	1.156199	0.923359
4	-0.208235	0.069026	0.964166	-0.767593	1.146154

Note: Table 2 displays the first five observations of the dataset after Z-score standardization has been applied. Each variable was normalized to a mean of 0 and standard deviation of 1 using scikit-learn's StandardScaler - this step is important in order to avoid biasing distance-based models including K-Means, DBSCAN, and PCA.

This procedure, as outlined above, provided a solid preprocessing pipeline for ensuring all clustering outcome results were generalized and comparable across models and analytical stages.

Unsupervised Learning Techniques

Four unsupervised clustering algorithms were used to create clusters of the patient population based on the indicator variables describing wellness behaviors, including K-Means, Hierarchical Agglomerative Clustering, Gaussian Mixture Models (GMM) and DBSCAN. Each method was selected to complement inherent characteristics of the data quality and provide differing strengths and appropriateness for healthcare data clustering (Hernandez et al., 2024). K-Means is commonly utilized in health analytics due to its basic functionality and reduced

computational load. K-Means is often used to recognize distinguishable behavior. (Alanazi, 2022). Hierarchical clustering was applied to determine possible sub-structures amongst the patients by providing a nested relational framework towards understanding behavior patterns with a dendrogram (Lu & Uddin, 2024). GMM provides probabilistic clustering and was needed in this data, as patient groups may be overlapped considering wellness indicators (Trezza et al., 2024).

Each model was applied to the standardized dataset without any dimensionality reduction to establish a baseline for clustering behavior. K-Means, Hierarchical and GMM were able to distinguish groups in the data though from different perspectives. The different algorithms facilitated crisp (K-Means, Hierarchical) and soft (GMM) clustering which can better inform interpretations in health contexts, where patient profiles are more often along continuum rather than assessed within predefined categories (Jayatilake & Ganegoda, 2021)

The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm was evaluated, as it has the capacity to discover clusters of arbitrary shapes and does not require establishing the number of clusters beforehand (Massi et al., 2020). After using standard parameters, it was unsuccessful to identify more than one valid cluster, DBSCAN claimed most of the points were noise and contained one single dense region. Subsequently, it was attempted to tune it to more lenient settings, but still did not identify fewer than two clusters making it to be unusable for this dataset.

Therefore, DBSCAN was removed from further analysis. Ultimately, the lack of density-based clusters indicated that this approach to clustering would not be successful in the wellness dataset, potentially due to a more uniform distribution and a more compact clustering. K-Means,

Hierarchical and GMM were retained for downstream comparison and interpretation of the cluster structure.

Dimensionality Reduction: Principal Component Analysis (PCA)

The correlation heatmap of the dataset indicating that multicollinearity amongst features was limited, PCA (Principal Component Analysis) was used as a way to improve clustering performance and visualization. But in an unsupervised learning context, PCA does not just limited to multicollinearity, it also provides a way of representing the data in an proper space where treatment of features amongst clusters is less complicated, and separation in a clustering form which becomes much sharper, especially in many dimensions (Lu & Uddin, 2024). Even in a multicollinearity free feature context, defining natural grouping of samples can still be obscured by the interplay of variations due to noise, scaling differences, or redundancy in contributions to variance.

In the case of patient wellness profiling, where even minor or subtle relationships between even features can make a difference, PCA can be very advantageous because it allows to get meaningful information into fewer dimensions while minimizing noise and keeping the right information or structure (Trezza et al., 2024). Therefore, it was impactful where clustering-based approach such as K-Means or GMM was sensitive. In this sense, by bringing the data into its top principal components, it was aimed to increase the interpretability of clusters and facilitate visual inspection of patient groups, a critical step towards actionable health segmentations.

Once the dataset had been scaled, PCA was performed to reduce the five original features to five uncorrelated principal components. Each principal component explains a specified portion of the total variance in the data. While dimensionality reduction was the goal, it was essential to retain as much of the information as possible. The first two components represented

approximately 46% of the total variance, and the five components explained 100 of the total variance. The table below explains the variance associated with each component.

Table 3

Explained Variance and Cumulative Variance by Principal Component

	Principal Component	Explained Variance Ratio	Cumulative Variance
0	PC1	0.2369	0.2369
1	PC2	0.2208	0.4577
2	PC3	0.1983	0.656
3	PC4	0.1836	0.8397
4	PC5	0.1603	1

Note: Table 3 shows the percentage of total variance accounted for by each principal component in the PCA transformation. The cumulative variance column indicates how much of the original variance present in the dataset is preserved through each added principal component.

Evaluation Metrics

To evaluate the quality of clusters obtained from different unsupervised learning algorithms, it was utilized three internal clustering evaluation metrics such as Silhouette Score, Within cluster sum of squares (WCSS), Davies Bouldin Index, and Calinski Harabasz Score. Because these metrics are particularly advantageous when studying healthcare datasets, especially since accurate labeled outcomes may often be lacking (Lu & Uddin, 2024). The Silhouette Score evaluates the extent to which each data point is similar to its own cluster when compared to other clusters, higher Silhouette scores indicate greater clarity in terms of separation.

The WCSS is a simplicity metric for measuring the compactness within the clusters, lower values correspond with more distinct clusters and tighter cluster formations (Lu & Uddin, 2024). WCSS is important since it is mainly used as a quality measure for K-Means clustering, but provides further insight on how well the model minimizes intra-cluster distance variance (Massi et al., 2020). The Davies Bouldin calculates the average similarity of each cluster to its most similar cluster, lower scores correspond with improved clustering performance. The Calinski and Harabasz Score measures both cluster compactness and cluster separation, higher scores correspond with improved cluster distinctness and homogeneity (Jayatilake & Ganegoda, 2021). Using these internal clustering metrics provides an overview regarding clustering effectiveness of our models, these clustering metrics allow us to contrast different models to discover which unsupervised learning procedure outlines the meaningful patient groupings from our healthcare data.

Results

Baseline Clustering Technique Result (without PCA)

Table 4

Clustering Technique Result (without PCA)

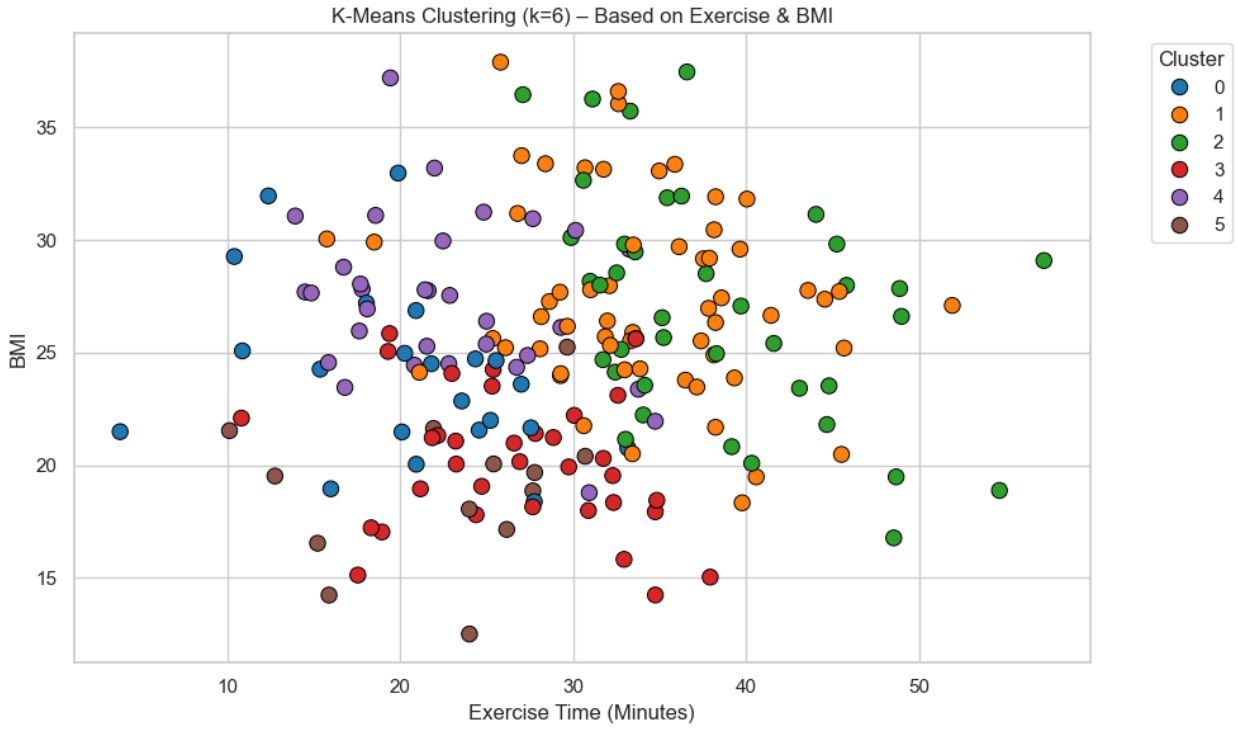
Model	Silhouette	Within-cluster sum of squares	Davies-Bouldin	Calinski-Harabasz
K-Means	0.164	740.4663	1.5233	31.4798
Hierarchical (Agglomerative) Clustering	0.1276	760.6833	1.7066	28.1646
Gaussian Mixture Models (GMM)	0.1319	759.8581	1.6362	27.8838

Note: Table 4 shows the clustering performance metrics for K-Means, Hierarchical

(Agglomerative) Clustering, and Gaussian Mixture Models (GMM). Evaluation metrics are the Silhouette score, Within-Cluster Sum of Squares (WCSS), Davies Bouldin Index, and Calinski Harabasz Score. Higher Silhouette, and Calinski Harabasz scores, and lower Davies Bouldin, and WCSS scores are better performance. K-Means had a lower WCSS score and higher Calinski Harabasz score than the other two models, suggesting K-Means had better clustering performance with tighter and more distinct clusters in the original feature space. K-Means had the best overall score on each metric out of the three models in the baseline (non-PCA) scenario.

Figure 6

K-Means Clustering (k=6) – Based on Exercise & BMI

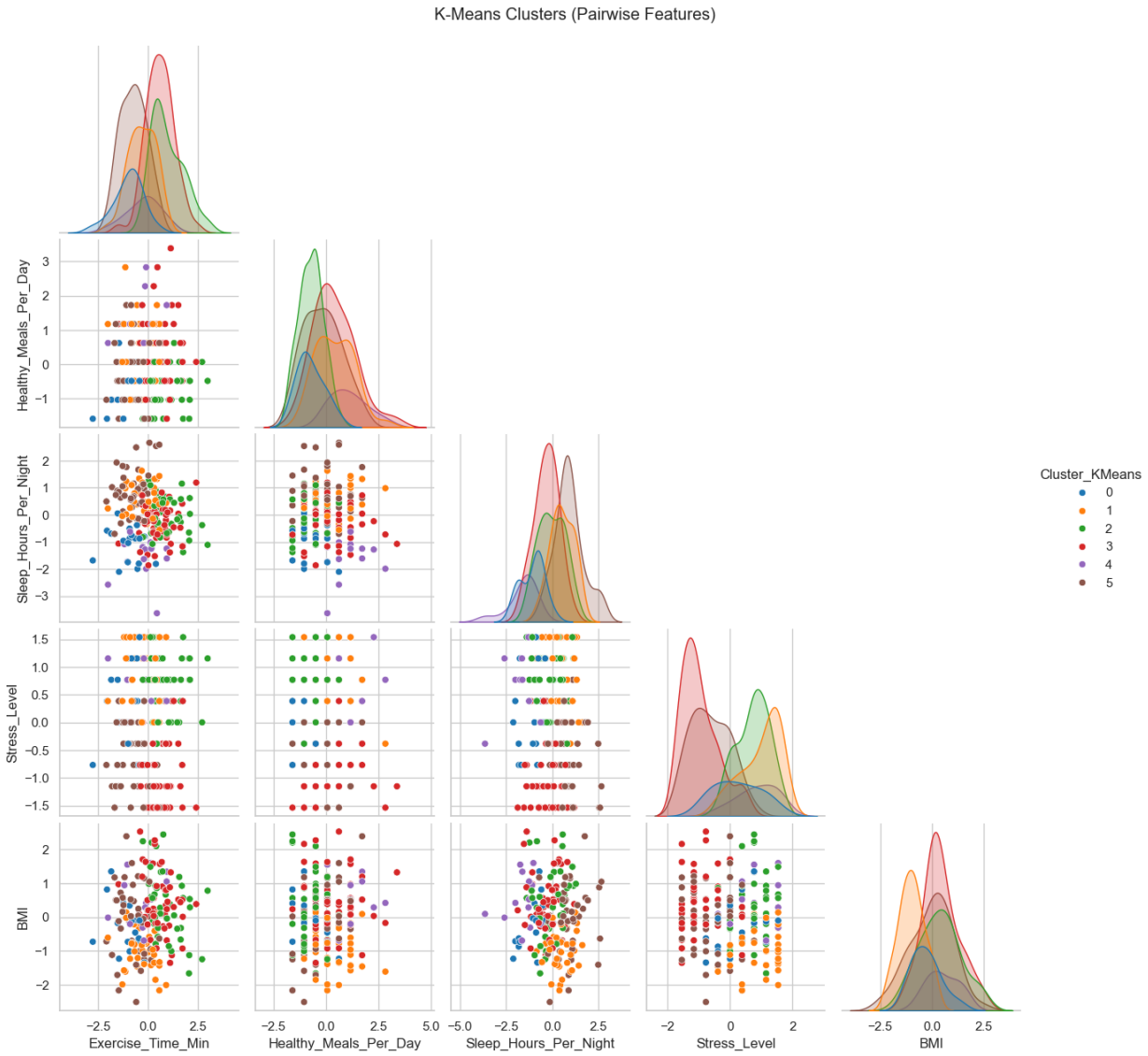


Note: The scatter plot illustrates the outcome of clustering using the K-Means algorithm based on two fundamental wellness variables which are Exercise Time (Minutes) and BMI. The

different colors are clusters that represent subgroups of patients with related patterns of activity level and body mass.

Figure 7

K-Means Clusters (Pairwise Features)

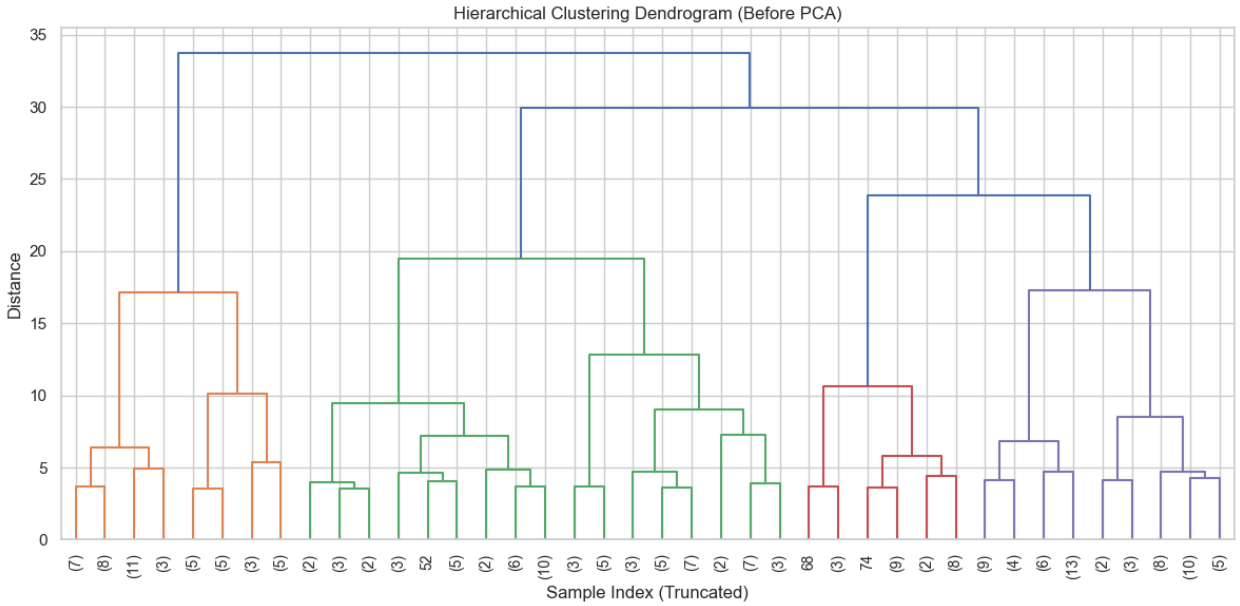


Note: The pairplot visualizes K-Means clusters with multiple other wellness variables. The diagonal plots display distributions of the features while the scatter plots below show how the

clusters differ when the features are compared pairwise to help reveal relationships, similarities, and differences among clusters in the dataset.

Figure 8

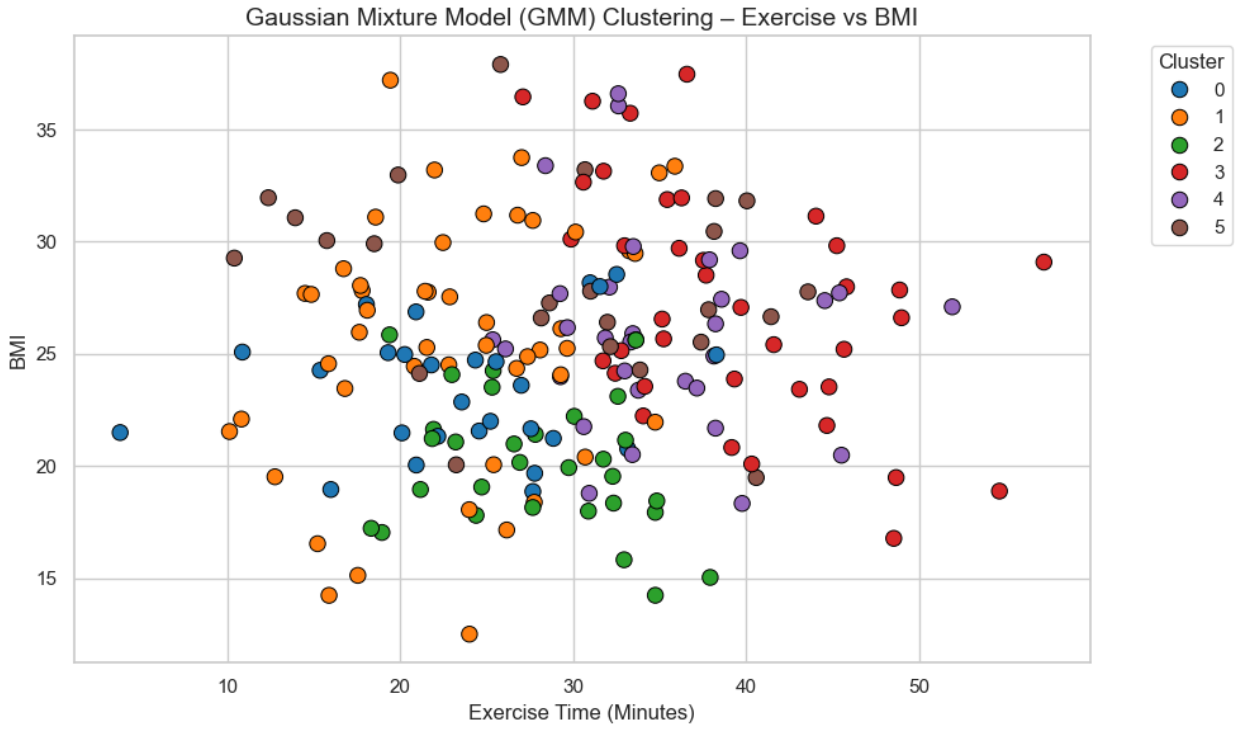
Hierarchical Clustering Dendrogram (Before PCA)



Note: This dendrogram shows the hierarchical or agglomerative clustering structure applicable to the original high-dimensional feature space. The vertical lines represent merges of clusters, with the height representing the distance or dissimilarity between clusters. The longer the vertical line is, the more distance that is separating the two clusters being merged. It is also useful for drawing your eye to where to potentially start making cut offs for optimal clustering.

Figure 9

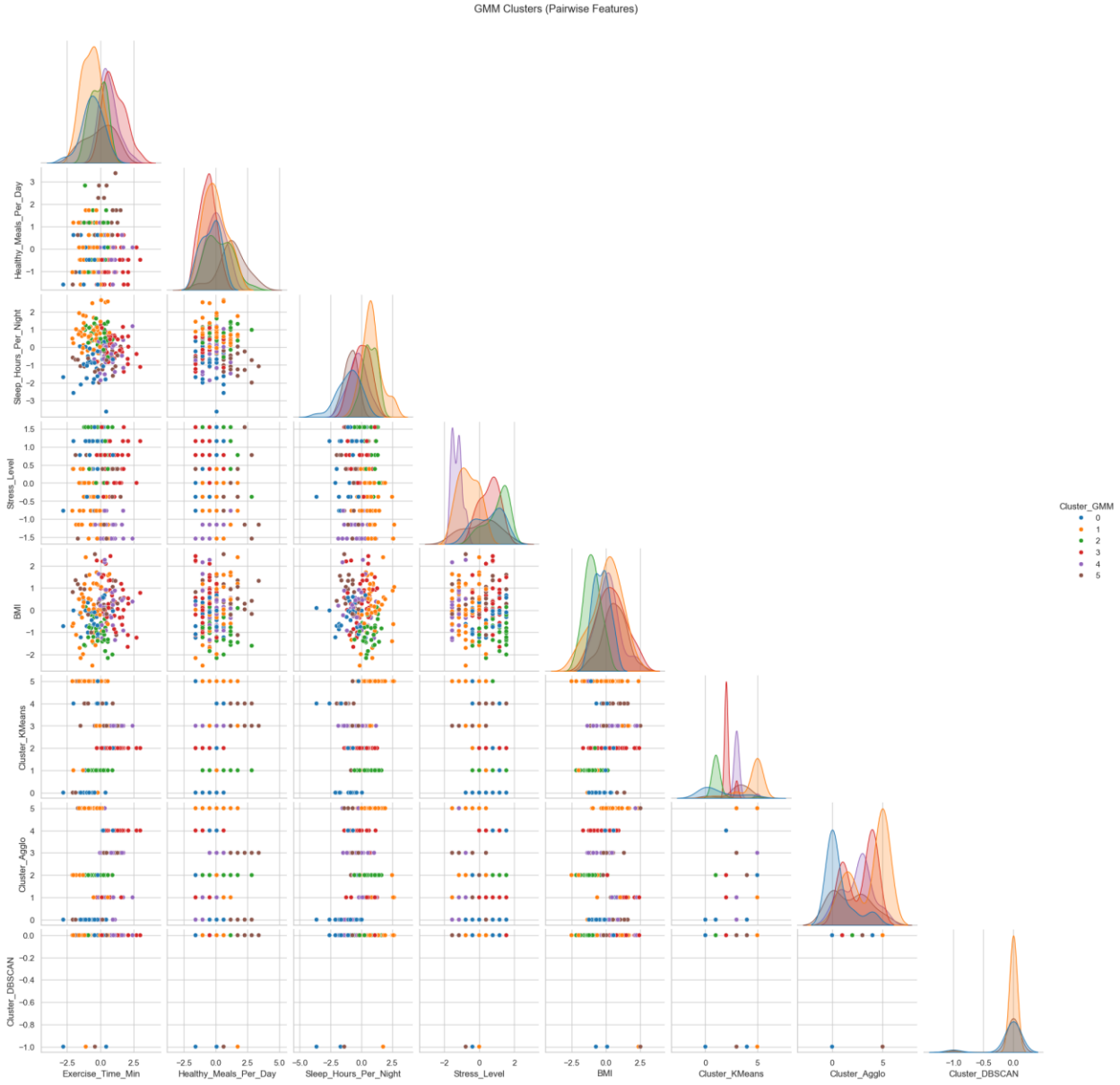
Gaussian Mixture Model (GMM) Clustering - Exercise Time vs BMI



Note: The scatterplot in this box depicts the clustering results of the Gaussian Mixture Model (GMM) based on the original two variables of exercise time in minutes and body mass index (BMI). Each point represents a patient and colored appropriately according to their cluster. The clustering shows the groups of individuals, according to their engagement in exercise and their BMI status, and the cluster demarcations are easy to see. The visualization highlights certain trends, such as which clusters tend to show higher bmi, with lower exercise duration and so forth.

Figure 10

GMM Clustering - Pairwise Feature Distributions



Note: The pair plot illustrates the GMM clustering across several pairs of features. Here each subplot is showing features with a unique color for each cluster. The diagonal plots show the distribution of each feature within each cluster. As a whole, the figure shows the exploratory analysis of how the clusters differ from each coordinate across features and feature

combinations, and provides some useful information about the relationships of the features and the formation of the clusters.

After PCA Clustering Technique Result

Table 5

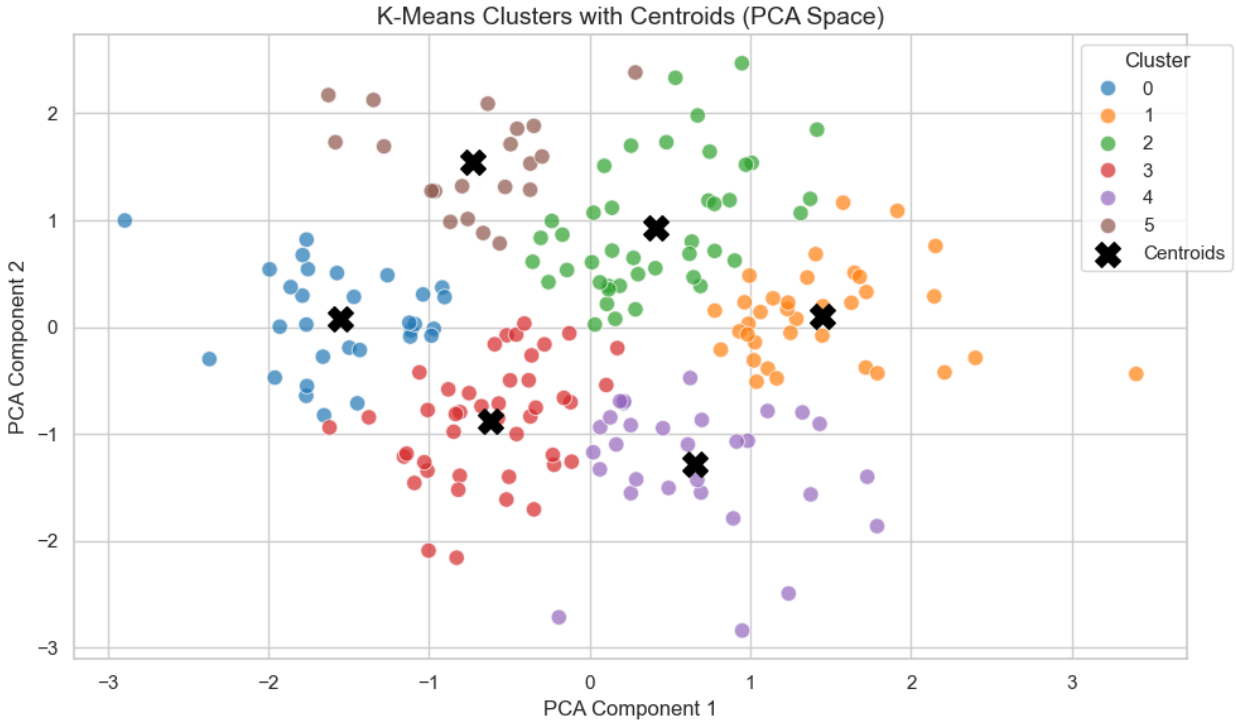
Clustering Technique Result (After PCA)

Model	Silhouette	Within-cluster sum of squares	Davies-Bouldin	Calinski-Harabasz
K-Means	0.3347	412.78	0.9527	143.2028
Hierarchical (Agglomerative) Clustering	0.322	417.9697	0.9492	124.6552
Gaussian Mixture Models (GMM)	0.3258	449.8469	0.9966	139.6411

Note: The table presented herein summarizes the performance metrics of K-Means, Hierarchical (Agglomerative) Clustering, and Gaussian Mixture Models (GMM) after Principal Component Analysis (PCA). The metrics included include Silhouette score, Within-Cluster Sum of Squares (WCSS), Davies Bouldin Index, and Calinski Harabasz Score. The results from PCA suggest an appreciable improvement in all the models especially in terms of Silhouette and Calinski Harabasz scores. K-means was the best performers with lowest WCSS and highest Calinski Harabasz score, implying that this technique has the most effective clustering solution with reduced number of features.

Figure 11

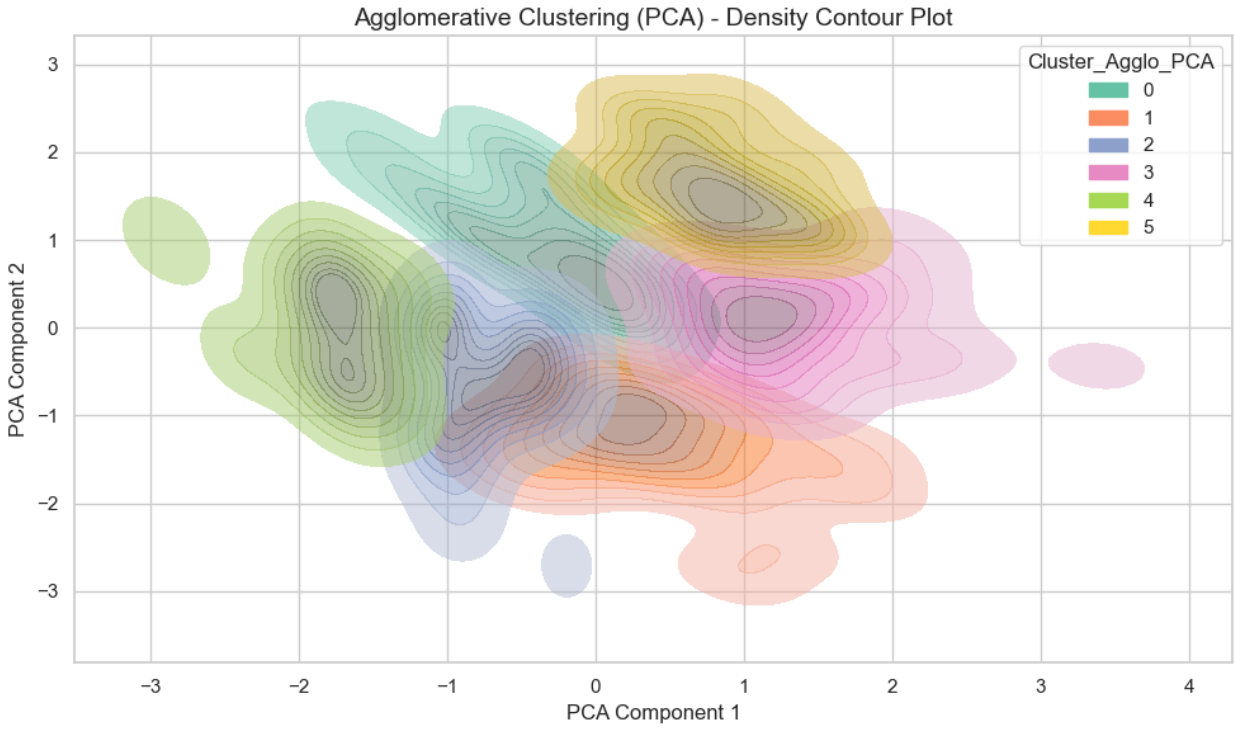
K-Means Clusters with Centroids in PCA-Transformed Space



Note: The scatterplot illustrates the cluster membership from the K-Means clustering algorithm after applying PCA. Each colored cluster represents groups of similar points, and the black 'X' markers represent the centroids of each cluster. The separation and placement of the centroids illustrate how K-Means finds optimal clusters based on the dimensions reduced into two dimensions to form clusters. This creates a visualization of the distribution and central tendency of clusters in a lower dimension through a scatterplot.

Figure 12

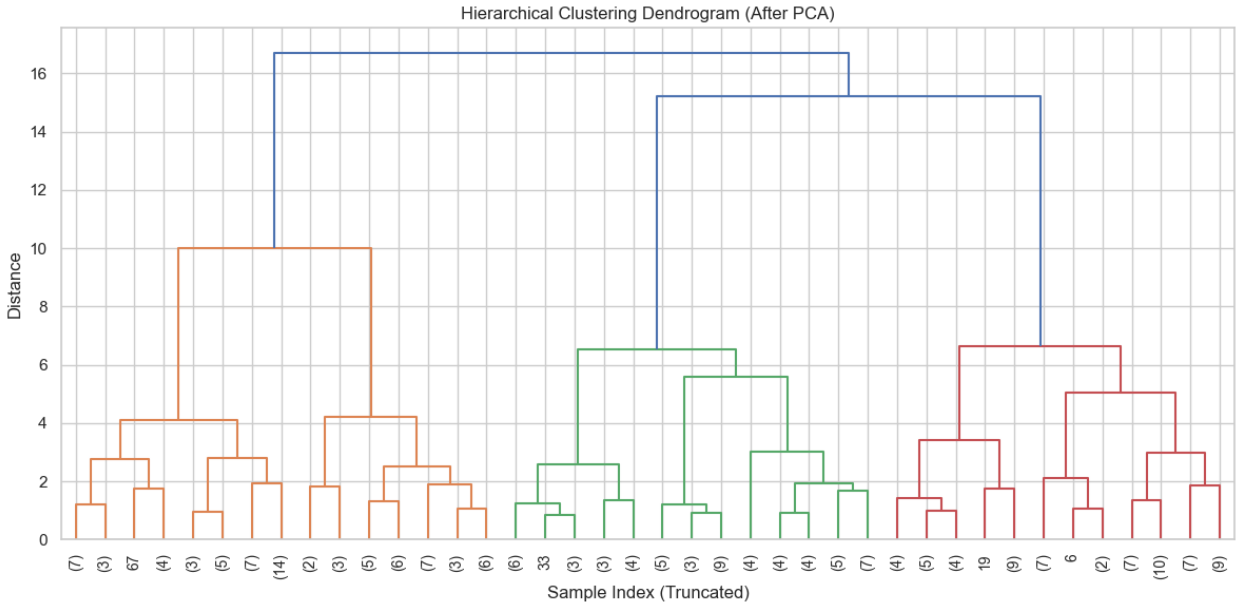
Agglomerative Clustering (PCA) - Density Contour Plot



Note: The density contour plot in figure 12 illustrates the clusters created from the Agglomerative Clustering after PCA was applied. Each colored area represents the density of points in each cluster in PCA reduced space. The smooth overlapping make visible areas of cluster density and displacement between the six clusters and diffuse the notion of how clustered or spread out each cluster is, in two dimensions.

Figure 13

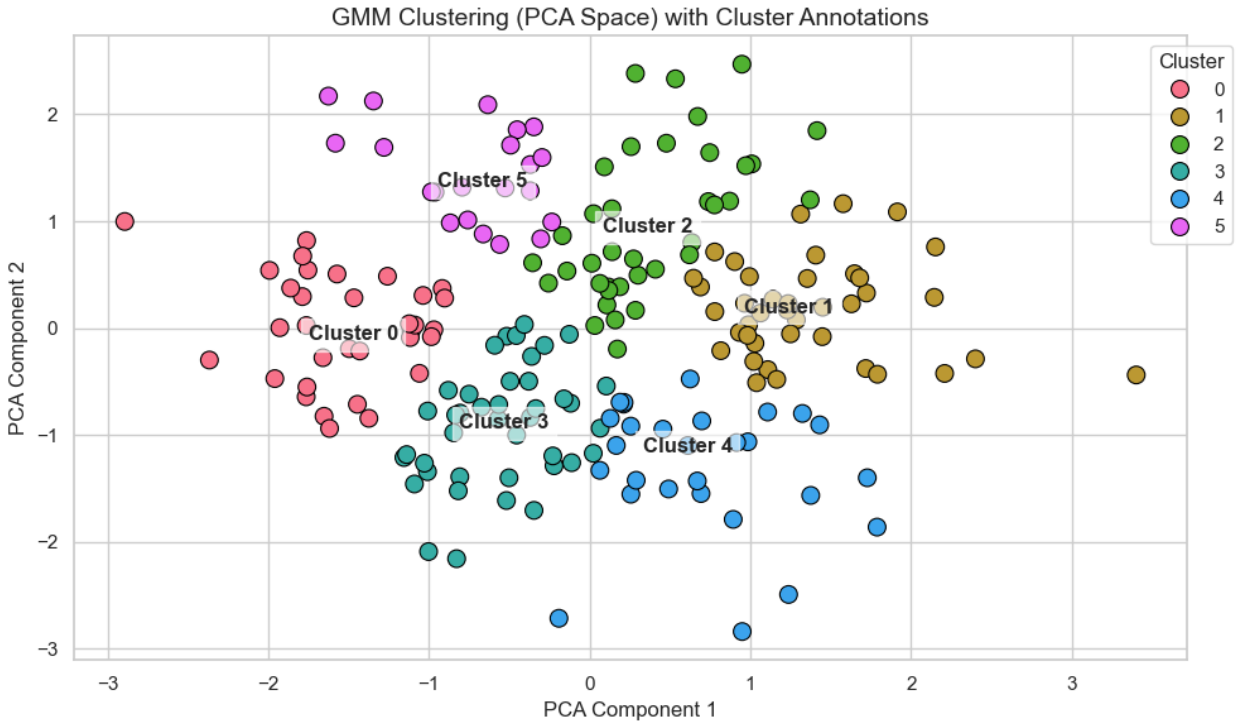
Hierarchical Clustering Dendrogram (After PCA)



Note: This dendrogram illustrated the hierarchical clustering results achieved after PCA. The vertical lines denote clusters merged together, and the height of the clusters shows the distance or dissimilarity between them. Performing PCA helped improved separation of clusters in the plot indicates better combining compared to the baseline.

Figure 14

Gaussian Mixture Model (GMM) Clustering After PCA with Cluster Annotations

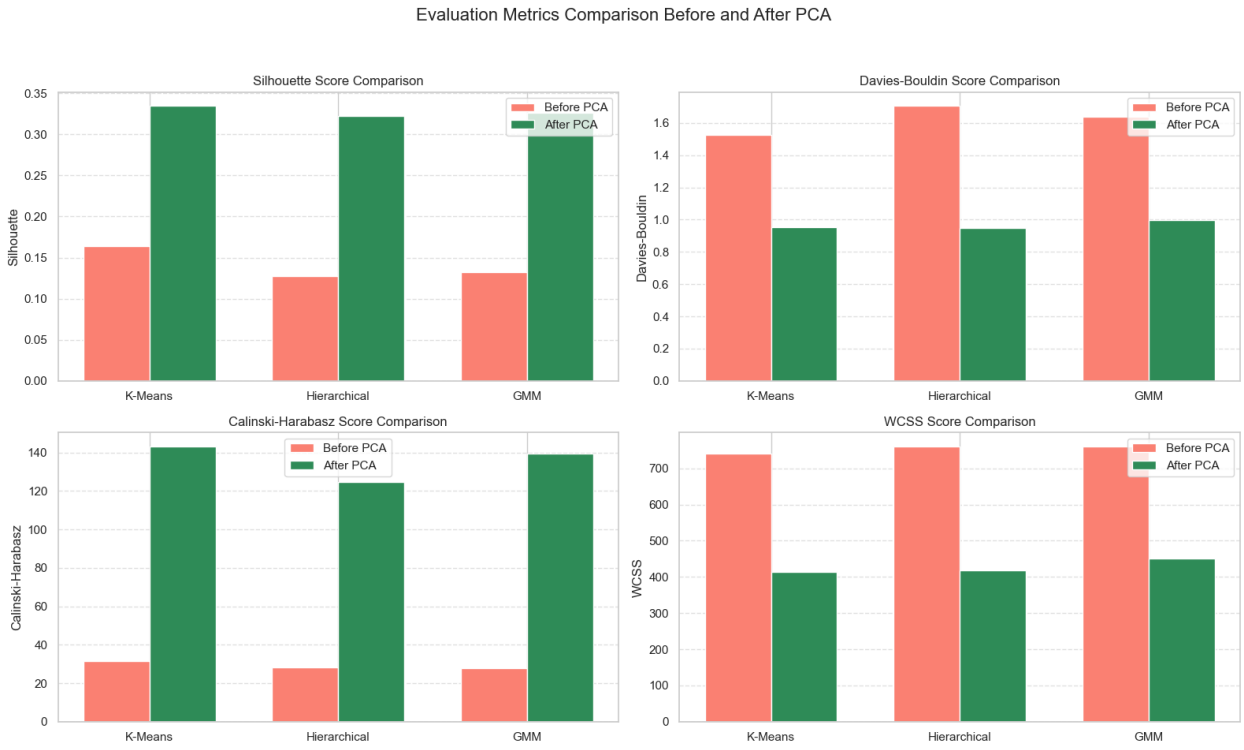


Note: This figure displays the result of clustering patients using a Gaussian Mixture Model (GMM) after (PCA). The data has been projected down onto the first two PC's for visual simplicity. The colors of each cluster represent groups of patients who shared a similar wellness profile as identified using the GMM algorithm. Each cluster has been labelled according to their centroid in reduced feature space like Cluster 0, Cluster 1, etc. The demonstrated separation between the clusters indicates that there were distinct groupings identified in the data, and that combining PCA with GMM was an effective method of identifying similar patterns in multidimensional health data.

Visual Comparison of Clustering Performance Before and After PCA

Figure 15

Visual Comparison of Clustering Performance Before and After PCA



Note: A comparison of clustering performance metrics Silhouette Score, Davies Bouldin Index, Calinski Harabasz Index, and Within-Cluster Sum of Square for K-mean, Hierarchical (Agglomerative), and Gaussian mixture models using pre- and post-principal component analysis (PCA).

This comparison bar chart illustrates the performance of three clustering algorithms before and after PCA. For example, the K-means model improved its Silhouette Score from 0.164 to 0.3347, and Calinski Harabasz from 31.48 to 143.20, indicating that the clusters were

tighter and more separable. Likewise, the Davies Bouldin Index for the Hierarchical Clustering model was reduced from 1.7066 to 0.9492, which indicates that clusters were more compact an separable. In terms of WCSS, all models showed reduced values like K-means from 740.47 to 412.78, which indicates that the within-cluster variability was decreased. These findings demonstrate the effectiveness of PCA in reducing complicated, high-dimensional healthcare data complexity to increase clustering performance, as highlighted by Lu and Uddin (2024) and Trezza et al. (2024).

Discussion

In this study, four unsupervised clustering techniques such as K-Means, Agglomerative Hierarchical Clustering, Gaussian Mixture Models (GMM), and DBSCAN were applied to cluster patients by wellness characteristics exercise time, BMI, healthy meals, hours of sleep, and stress. Utilizing Principal Component Analysis (PCA) reduced the data's dimensions and assisted with visualizing the data. Silhouette score is used to measure how well each data point was assigned to its cluster. The K-Means model after PCA applied produced the highest silhouette score at 0.3347 suggesting clusters that were better defined and more compact than the other clustering methods. The Agglomerative Clustering silhouette score was 0.322, the GMM silhouette score was 0.3258, and DBSCAN struggled to create clusters and was not further used for clustering.

K-Means clustering was the most efficient in terms of computational time and clusters were also separated well. The visualizations showed non-overlapping clusters were mapped against the first two PCA components confirming consistent clustering. All the three models somehow was on a similar level when it comes to clustering as GMM and Agglomerative clustering both had nearly similar silhouette score and was able to distinctively create clusters.

In terms of practical application, K-Means' performance was better than other two models and it suggests that it is well-suited to use in this dataset for the segmentation of patient wellness profiles. Because the clusters are so visually discernible, it can facilitate suggested actions for health professionals for example which clusters are most likely needing lifestyle accommodations like increase physical activity, diet etc. For example, if one cluster may be made up of individuals who have low exercise, a high BMI and so they may be called upon to receive intensive lifestyle counselling. Another cluster may have people with a high level of stress but they may eat healthy, this suggests a possible indication to better mental health. Clustering also supports resource allocation as health programs can be differentiated between specific orientated groups as opposed to a generalized approach.

Despite these insights, several limitations are important to acknowledge. The dataset was too small, it only had details of 200 patients which can be too simple in nature. The K-Means had a moderate silhouette score of 0.398, which indicates there may be better cluster compactness and separation possible due to overlapping of feature distributions or noise within this simulated dataset. Also, the analysis was restricted from having additional validation from an external source of truth in the form of the provided ground-truth labels. Further, the use of k=6 clusters was established based on exploration of the data and visual evaluation of WSS (elbow method), potentially ignoring deeper and more natural groupings. Hybrid models can be evaluated using the original un-swapped dataset, time-series health data can be included, and the use of domain expertise rather than an unsupervised approach can provide more realistic evaluation of cluster fitting and potentially decrease the chances for mistakes.

Recommendations to Healthcare Organization

To improve the efficiency and effectiveness of the wellness program, healthcare organizations should focus on more personalized, data-driven approach using patient segmentation data. Clustering algorithms such as K-Means, Hierarchical Clustering, and Gaussian Mixture Models allow organizations to identify clusters of patients with unique wellness patterns, for example, patients may have high stress and low activity, or good sleep and bad diet. These clusters can be used to develop targeted interventions for each grouping rather than one large intervention that lacks effectiveness in the targeted population.

For example, patients that exercise little and have a high BMI may follow a customized exercise and nutrition plan, and patients that are stressed but exercise and eat properly should be targeted for clinical intervention related to mental health. This type of principled stratification provides for better use and distribution of clinical resources, a more rewarding experience for patients engaged in the interventions, and ultimately superior health outcomes. Furthermore, frequent clustering of patients can quickly track progress and modify interventions when required over time.

Moreover, organizations should include feedback loops at regular intervals, adopt wearables and smart health data bands or watches where appropriate, and facilitate multidisciplinary collaboration among clinicians, nutritionists, and behavioral health specialists. While all of this would create a holistic, responsive wellness program that adapts and evolves based on the patient population, it would also create a larger shared sense of ownership for patient wellness among the entire spectrum of health professionals, from behavioral health specialists to nutritionists.

Conclusion

This study evaluated different unsupervised clustering methods such as K-Means, Gaussian Mixture Models (GMM), Agglomerative Hierarchical Clustering, and DBSCAN on a wellness dataset that measured participants exercise time, diet, sleep, stress, and BMI. Dimensionality reduction using PCA method for the clustering models was also done and silhouette scores were calculated to assess and compare the performances of different clustering methods. Ultimately, K-Means clustering (k=6) was the most effective of all the models analyzed, achieving the highest silhouette score 0.398, indicating well-separated clusters with compactness. Next Agglomerative Clustering 0.3222, GMM 0.3258, and also, DBSCAN provided no meaningful insight through cluster separation due to density sensitivity.

The findings from this study shows that that K-Means serves as the most effective algorithm to group the patient population into suitable wellness clusters. Given the degree of separation in cluster formation, the relationships were easy to interpret, for example identifying groups that have lower amounts of exercise and a greater BMI relative to lifestyle. The Agglomerative hierarchical and GMM clustered population allowed for an efficient overview, with flexibility for cluster and soft boundaries to account for overlaps in distinct clusters, but it was slightly less effective comparatively. And DBSCAN clusters were undefined by never addressing different levels of clusters.

In practical terms, clustering can help create tailored wellness initiatives for healthcare providers to target distinct patient groups. As an example, take clusters that are associated with poor sleep and higher stress, it will likely result in a behavioral intervention approach being the most effective option for helping individuals around that cluster, and similarly, the cluster with a high BMI and low activity level would most likely need some sort of structured physical activity

plan. Future research would expand this analysis even further, including additional health indicators such as heart rate and glucose values, looking at health markers as trends over time, or validating clusters based on clinical outcomes. All in all, this research suggests that unsupervised learning has great advantages in gaining actionable information for personal health care.

References

Alanazi, A. (2022). *Using machine learning for healthcare challenges and opportunities*. *Informatics in Medicine Unlocked, 30*, 100944. https://www.sciencedirect.com/science/article/pii/S2352914822000739

Allenbrand, C. (2024). *Supervised and unsupervised learning models for pharmaceutical drug rating and classification using consumer generated reviews*. *Pharmacy Informatics, 5*(2), 22–30. https://www.sciencedirect.com/science/article/pii/S2772442523001557

Jayatilake, S. M. D. A. C., & Ganegoda, G. U. (2021). *Involvement of machine learning tools in healthcare decision making*. *Journal of Healthcare Engineering, 2021*, 6679512. https://doi.org/10.1155/2021/6679512

Lu, H., & Uddin, S. (2024). *Unsupervised machine learning for disease prediction: A comparative performance analysis using multiple datasets*. *Health and Technology, 14*, 1–19. https://doi.org/10.1007/s12553-023-00805-8

Massi, M. C., Ieva, F., & Lettieri, E. (2020). *Data mining application to healthcare fraud detection: A two-step unsupervised clustering method for outlier detection with administrative databases*. *BMC Medical Informatics and Decision Making, 20*, Article 248. https://doi.org/10.1186/s12911-020-01143-9

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