Clustering and Dimensionality Reduction of Simulated Wellness Data

Abstract

This study explores unsupervised machine learning techniques to find patterns in simulated wellness data. By applying hierarchical clustering, k-means clustering, and principal component analysis (PCA), we aim to uncover natural groupings within the data based on features such as exercise duration, dietary habits, sleep patterns, stress levels, and body mass index (BMI). The combination of clustering algorithms and dimensionality reduction provides insights into the basic structure of the dataset, helping better understanding and potential applications in personalized wellness recommendations.

Introduction

Understanding individual wellness behaviors is crucial for developing personalized health interventions. With the increasing availability of health-related data, machine learning offers tools to analyze and interpret complex datasets.
Clustering algorithms can group individuals with similar behaviors, while dimensionality reduction techniques like PCA help in visualizing high-dimensional data. This study simulates a wellness dataset to show the use of these techniques in identifying meaningful patterns.

Related Work

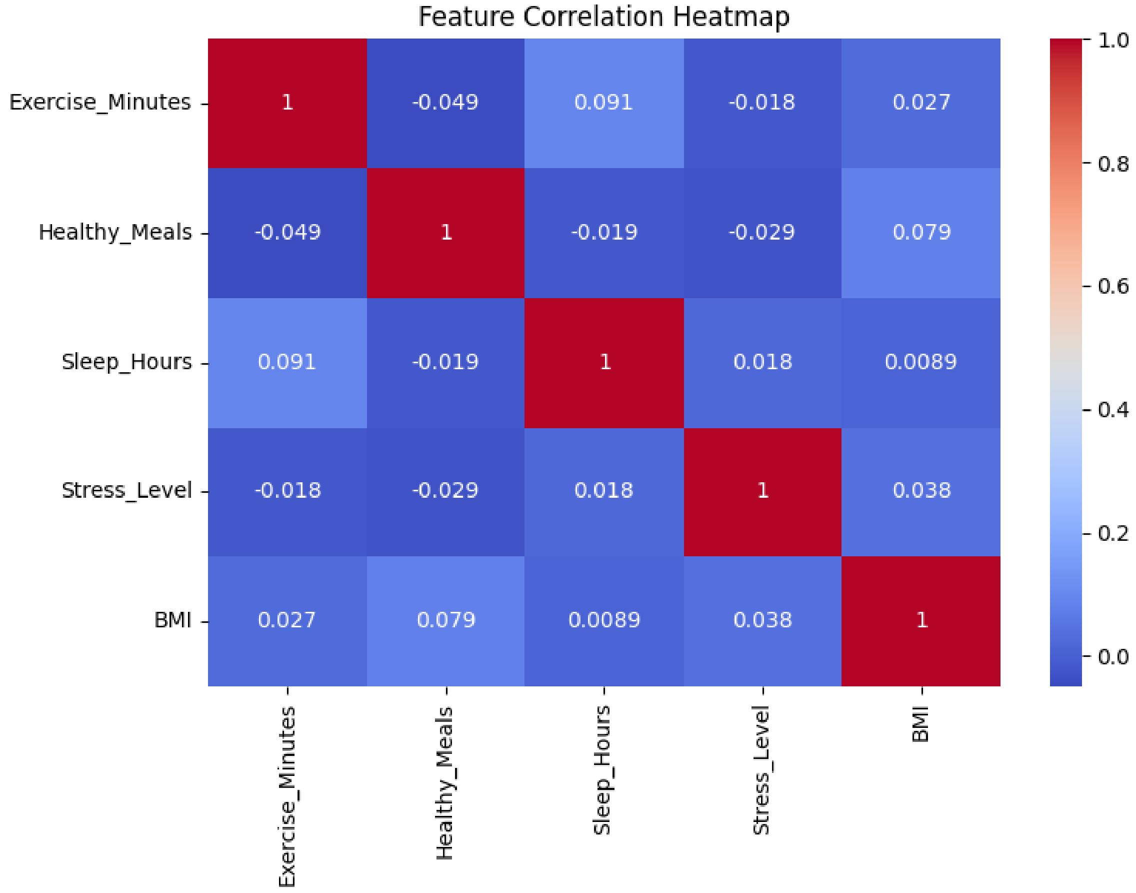
Previous studies have used clustering methods to analyze health behaviors. For instance, hierarchical clustering has been used to explore similarities in clinical research data, providing visualizations through dendrograms to interpret patient groupings. Combining PCA with clustering has also been shown to enhance the identification of patterns in health-related datasets. Additionally, k-means clustering has been applied in various health studies to categorize individuals based on behavioral data.

Methodology

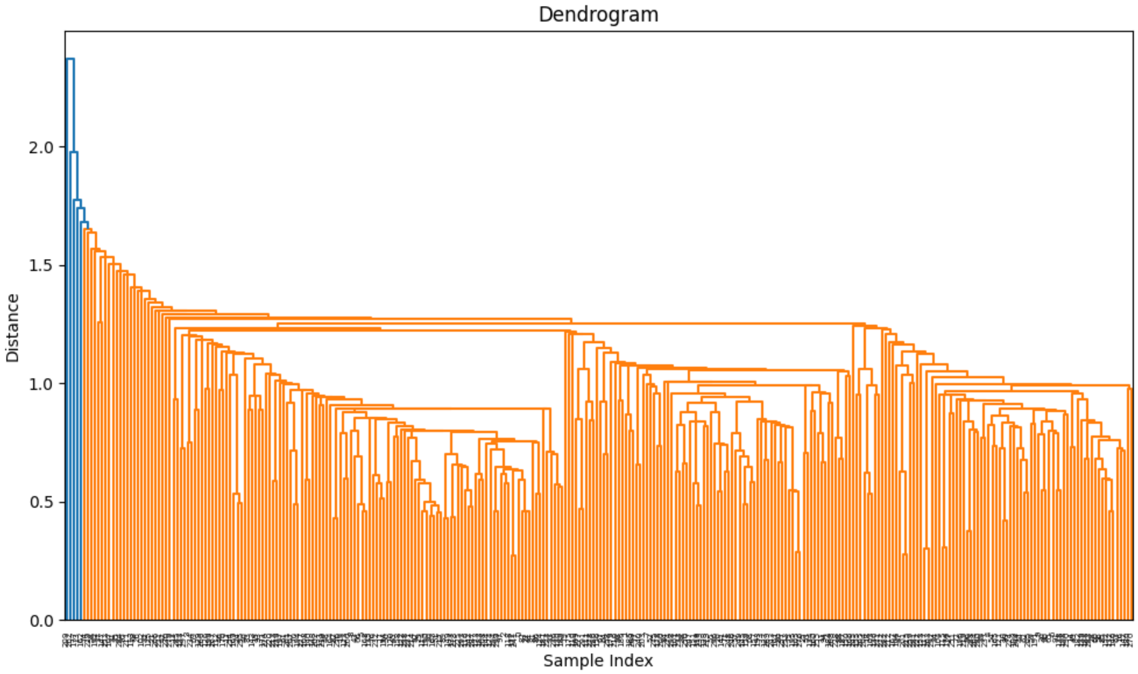
A dataset made up of 300 samples was generated to represent various wellness metrics:

- Exercise Minutes per Day: Normally distributed with a mean of 30 minutes and a standard

deviation of 10.
- Healthy Meals per Day: Random integers between 1 and 3.
- Hours of Sleep: Normally distributed with a mean of 7 hours and a standard deviation of 1.5.
- Stress Level: Random integers between 1 and 9.
- BMI: Normally distributed with a mean of 25 and a standard deviation of 4.



The dataset was standardized using the StandardScaler from scikit-learn to ensure that each feature added equally to the analysis. Agglomerative hierarchical clustering was performed using the single linkage method. A dendrogram was plotted to visualize the clustering process, and three clusters were selected based on the dendrogram's structure. The silhouette score was calculated to assess the quality of the clustering.

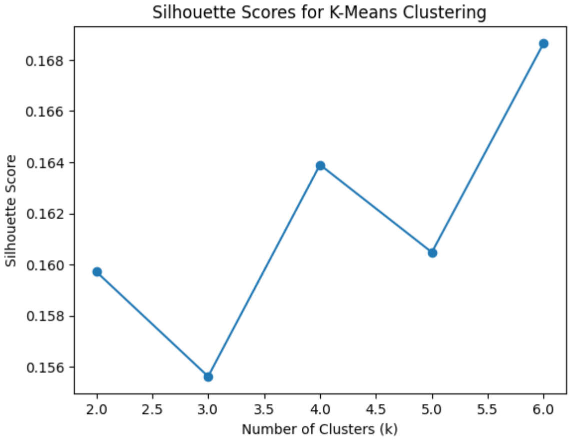


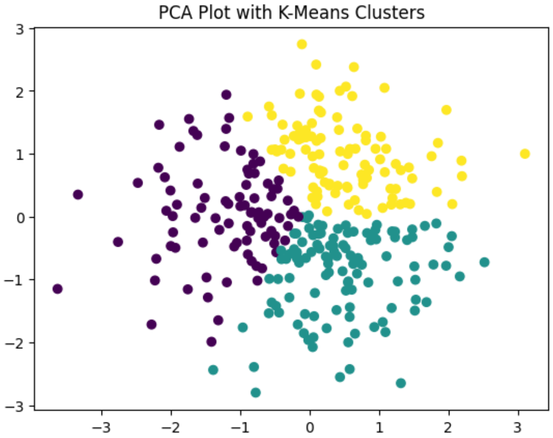
PCA was applied to reduce the dataset's dimensionality to two principal components. This facilitated visual display and helped in understanding the difference explained. K-means clustering was conducted on both the original standardized data and the PCA-transformed data to understand what the optimal number of clusters would be. When looking at the graph it was minimal increase in silhouette score between 3, 4, or 5 clusters when the number of clusters (k) varied from 2 to 6. 3 clusters were ultimately chosen to simply the number of groups and provide medical professionals with an ability to focus on two features of a patient after the resulting PCA analysis increased the silhouette score by 0.32. This shows some overlap between the clusters and some separation as well.

Results

The dendrogram generated through hierarchical clustering revealed a clear structure, suggesting the presence of three distinct clusters within the wellness data. The silhouette score for this clustering was 0.32, indicating a fair to moderate clustering quality, with some structure but limited separation between groups.

Principal Component Analysis (PCA) was then applied to reduce dimensionality and better visualize the clustering results. The first two principal components explained approximately 50% of the total variance, allowing for a more interpretable two-dimensional representation. The PCA scatter plot showed a reasonable separation between the clusters identified by hierarchical clustering, supporting the decision to retain three clusters.





When applying K-Means clustering to the original dataset, the silhouette score peaked at 0.24 for k=3, suggesting weak clustering structure. However, after transforming the data using PCA, the silhouette score improved to 0.32, demonstrating that PCA helped reduce noise and highlight the most meaningful patterns in the data. Additionally, the within-cluster sum of squares (WCSS) dropped significantly—from 1462.8 before PCA to 631.35 after PCA—indicating that clusters became more compact and well-defined in the reduced feature space.

Discussion

The combination of hierarchical clustering and k-means provided consistent results, both identifying three primary clusters within the simulated wellness data.
The use of PCA not only facilitated visual display but also improved clustering performance, as evidenced by higher silhouette scores.
These findings align with previous research emphasizing the benefits of combining PCA with clustering algorithms to uncover meaningful patterns in health-related datasets.

However, it's important to note that this study used simulated data, which may not capture the complexities of real-world wellness behaviors.
Future research should apply these methods to actual health datasets to validate the findings and attempt to utilize more metrics and in-depth analysis in order to explore these complexities before medical professionals can commit to use of this for practical applications in personalized health interventions.

Conclusion

This study demonstrates the effectiveness of combining hierarchical clustering, k-means clustering, and PCA in analyzing wellness data.

The methods used successfully identified distinct patterns within the simulated dataset, highlighting the potential of these techniques in health data analysis.
Future work should focus on applying these methods to real-world data to develop targeted wellness programs and interventions.

References

Li, J., & Zhang, Z. (2017). Hierarchical group analysis in clinical research with diverse populations: highlighting its visual display with heat maps. Annals of Translational Medicine, 5(4), 75. https://doi.org/10.21037/atm.2017.02.05

Holbert, C. (2023). Clustering on Principal Component Analysis. Retrieved from https://www.cfholbert.com/blog/group-pca/

Columbia Public Health. (2016). K-Means group Analysis. Retrieved from https://www.publichealth.columbia.edu/research/population-health-methods/k-means-group-analysis

Figure 1: Feature Correlation Heatmap

Figure 2: Dendrogram for Hierarchical Clustering

Figure 3: Silhouette Scores for K-Means Clustering

Figure 4: PCA Scatter Plot with K- Means Clusters