Unsupervised Learning for Health Cluster Analysis: Investigating the Impact of Socioeconomic Factors on Health Clusters in NHANES Data

John Vernon Baldeo
College of Computing and Information Technology(CCIT) National University(N.U.)
Manila, Philippines
baldeojb@students.national-u.edu.ph

Gian Karl Colinares
College of Computing and Information Technology(CCIT) National University(N.U.)
Manila, Philippines
colinaresgl@students.national-u.edu.ph

Carl Arvin Hipolito
College of Computing and Information Technology(CCIT) National University(N.U.)
Manila, Philippines
hipolitocc@students.national-u.edu.ph

***Abstract*—This study explores the relationship between socioeconomic status (SES) and health outcomes using unsupervised learning on NHANES data. K-Means and DBSCAN clustering were applied to key SES factors (education, family poverty ratio, housing) and health indicators (BMI, blood pressure, CRP, glycohemoglobin). Clusters were analyzed in both raw and PCA-transformed feature spaces, revealing distinct health risk patterns. Results indicate that lower SES groups exhibit higher metabolic risks, while those with better healthcare access demonstrate improved outcomes. These findings highlight the potential of data-driven clustering to inform targeted public health interventions and reduce health disparities.**

***Index Terms*—Unsupervised Learning, Clustering Algorithms, K-Means, DBSCAN, Agglomerative Clustering, Socioeconomic Status (SES), Health Disparities, Healthcare Access, Morbidity Score, NHANES Data Analysis.**

I. Introduction

Understanding the relationship between socioeconomic status (SES), demographic factors, and health outcomes is a critical area of research in public health and epidemiology. Socioeconomic disparities significantly influence health risk factors, disease prevalence, and access to healthcare services, ultimately shaping population health trends [1]. Traditional statistical methods have provided valuable insights into these relationships; however, they often fail to capture complex, nonlinear patterns in multidimensional health data. With the increasing availability of large-scale health datasets such as the National Health and Nutrition Examination Survey (NHANES), machine learning approaches—particularly unsupervised learning techniques—offer an opportunity to uncover hidden health patterns without predefined assumptions [2].

Unsupervised learning algorithms, such as k-means clustering, hierarchical clustering, and DBSCAN, allow researchers to identify latent health clusters based on a combination of physiological, behavioral, and socioeconomic variables. These clusters can help in stratifying populations based on health risk profiles, enabling targeted interventions and resource allocation [3]. Previous studies have applied clustering methods to NHANES data to segment individuals based on dietary habits, physical activity, and chronic disease risk factors, yet there remains a gap in exploring how SES and demographic factors influence these health clusters [4].

This study aims to apply unsupervised machine learning techniques to NHANES data to investigate how SES and demographic factors contribute to distinct health profiles. By integrating clustering algorithms with key health indicators, including BMI, blood pressure, cholesterol levels, and diabetes markers, this research seeks to identify patterns of health disparities linked to income, education, and poverty. The findings could provide a data-driven approach for policymakers and healthcare professionals to design more targeted interventions for vulnerable populations.

II. Review of Related Literature

A. *Socioeconomic Determinants of Health*

Socioeconomic status (SES) plays a crucial role in shaping health outcomes, influencing both the prevalence and severity of chronic diseases. Research has consistently demonstrated that individuals with lower SES are at a higher risk of developing conditions such as obesity, cardiovascular disease, and diabetes due to limited access to healthcare, nutritious food, and healthy living conditions [5]. Wilkinson and Marmot [6] argue that socioeconomic inequalities directly correlate with disparities in morbidity and mortality rates, with income and education serving as major predictors of health status.

Furthermore, the World Health Organization (WHO) highlights that social determinants of health—including financial stability, educational attainment, and employment—affect lifestyle choices, stress levels, and healthcare utilization, which collectively contribute to long-term health risks [7]. Recent studies using NHANES data have found that lower-income individuals tend to exhibit higher levels of obesity and hypertension due to poor dietary habits and reduced physical activity [8]. Additionally, access to medical care and preventive screenings remains disproportionately limited for individuals in lower-income brackets, exacerbating health disparities [9].

B. *Clustering Approaches in Health Data*

Unsupervised learning techniques, particularly clustering algorithms, have been widely utilized in healthcare analytics to identify hidden patterns in patient populations. Clustering methods such as k-means, hierarchical clustering, and density-based spatial clustering (DBSCAN) have proven effective in segmenting individuals based on health profiles [10]. These methods enable the identification of groups with similar risk factors, allowing for targeted interventions and more personalized healthcare approaches [11].

For example, a study by Smith et al. [8] employed hierarchical clustering to analyze NHANES data and successfully categorized individuals into distinct health clusters based on their metabolic risk factors. The study revealed significant associations between SES and cluster membership, with lower-income participants more likely to be in high-risk clusters characterized by obesity and hypertension. Another research by Kim et al. [9] demonstrated the efficacy of DBSCAN in identifying outliers in NHANES health data, which helped detect rare conditions that traditional classification methods often overlook.

Clustering has also been used to segment dietary patterns and physical activity behaviors, highlighting how socioeconomic disparities influence health behaviors. Studies have found that individuals in lower-income brackets tend to cluster in dietary groups characterized by high sugar intake and low nutritional value, leading to increased risks of chronic diseases [14]. Such findings reinforce the need for integrating SES variables in clustering models to better understand health disparities.

C. *Machine Learning Applications in NHANES Data*

The increasing availability of large-scale health datasets such as NHANES has facilitated the adoption of machine learning techniques for public health research. Supervised and unsupervised learning methods have been widely applied to analyze disease risk factors, predict health outcomes, and uncover hidden patterns within population health data [15].

In a study conducted by Patel et al. [16], k-means clustering was used to classify NHANES participants into distinct health groups based on biomarker data, revealing that individuals in low-income communities exhibited a significantly higher prevalence of metabolic syndrome. Similarly, Wang et al. [11] applied an ensemble clustering approach to examine the relationship between physical activity levels and cardiovascular disease risk. Their findings emphasized the need for targeted health interventions tailored to socioeconomically disadvantaged populations.

Another study by Larson and Brown [14] utilized deep learning-based clustering techniques to analyze NHANES biometric and lifestyle data, identifying patterns associated with comorbidities such as obesity, diabetes, and hypertension. The integration of SES variables in these clustering models allowed for the identification of high-risk groups that traditional epidemiological approaches often fail to capture.

Despite these advancements, there remains a gap in research exploring how SES and demographic factors influence health clusters in NHANES data. Existing studies focus primarily on disease prediction or individual health behaviors, but few have systematically examined how clustering techniques can be applied to uncover disparities in health outcomes based on socioeconomic and demographic indicators. This study seeks to address this gap by leveraging unsupervised learning methods to analyze NHANES data, with a particular emphasis on SES-driven health patterns.

III. Clustering Methods

Clustering is a fundamental data analysis technique used in pattern recognition, image segmentation, anomaly detection, and various other applications. It partitions a dataset into clusters, ensuring that intra-cluster similarity is maximized while inter-cluster similarity is minimized. Various clustering algorithms exist, each employing distinct principles for data partitioning. This paper discusses three widely used clustering techniques: K-Means, DBSCAN, and Agglomerative Hierarchical Clustering.

A. *K-Means Clustering*

K-Means is a widely used centroid-based clustering algorithm that aims to minimize the variance within each cluster by iteratively updating cluster centroids. It is computationally efficient and effective for well-separated, spherical clusters[17].

*1) **Mathematical Formulation**:* Given a dataset with n data points, the K-Means algorithm partitions them into k clusters by minimizing the intra-cluster variance[18]:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2, \tag{1}$$

where C_i represents cluster i, μ_i is the centroid of cluster i, and $||x - \mu_i||^2$ is the squared Euclidean distance.

*2) **Parameters**:*

- k: Number of clusters.
- Centroids: Initial cluster centers (randomly selected or initialized via K-Means++).
- Stopping criterion: Maximum iterations or convergence threshold.

*3) **Performance Measures**:*

- Sum of Squared Errors (SSE): Measures the total squared distance of data points from their respective centroids.
- Inertia: The total within-cluster sum of squared distances.
- Silhouette Score: Evaluates how well each data point fits within its assigned cluster.
- Davies-Bouldin Index (DBI): Measures cluster compactness and separation.
- Calinski-Harabasz Index: Assesses cluster dispersion and separation.

B. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN is a density-based clustering algorithm that groups closely packed points in dense regions while marking outliers as noise. Unlike K-Means, DBSCAN does not require the number of clusters to be specified in advance and can detect clusters of arbitrary shape[19].

*1) **Mathematical Formulation**:* DBSCAN relies on two parameters: ε (neighborhood radius) and $minPts$ (minimum points in a cluster). A point p is a core point if:

$$|N_\varepsilon(p)| \geq minPts, \tag{2}$$

where $N_\varepsilon(p)$ represents the set of points within distance ε of p [20].

*2) **Parameters**:*

- ε: Neighborhood radius.
- $minPts$: Minimum points required to form a dense region.

*3) **Performance Measures**:*

- Silhouette Score
- Davies-Bouldin Index (DBI)
- Adjusted Rand Index (ARI)

C. Agglomerative Hierarchical Clustering

Agglomerative clustering is a bottom-up hierarchical clustering technique that iteratively merges clusters based on a linkage criterion [21].

*1) **Mathematical Formulation**:* Given two clusters A and B, the linkage function $d(A, B)$ can be defined as:

$$d(A, B) = \min_{a \in A, b \in B} d(a, b) \quad \text{(Single Linkage)}, \tag{3}$$

$$d(A, B) = \max_{a \in A, b \in B} d(a, b) \quad \text{(Complete Linkage)}, \tag{4}$$

$$d(A, B) = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a, b) \quad \text{(Average Linkage)}. \tag{5}$$

*2) **Parameters**:*

- Linkage criterion: Single, complete, or average linkage.
- Distance metric: Euclidean, Manhattan, etc.
- Number of clusters (optional if cutting the dendrogram).

*3) **Performance Measures**:*

- Silhouette Score
- Calinski-Harabasz Index
- Davies-Bouldin Index (DBI)

D. Performance Measures

To evaluate the effectiveness of clustering algorithms, several internal validation metrics are used[20].

*1) **Silhouette Score**:* The Silhouette Score measures how similar a data point is to its own cluster compared to other clusters. It is defined as:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}, \tag{6}$$

where $a(i)$ is the average intra-cluster distance, and $b(i)$ is the lowest average inter-cluster distance.

*2) **Davies-Bouldin Index (DBI)**:* DBI evaluates clustering quality based on the compactness and separation of clusters:

$$DBI = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \frac{s_i + s_j}{d_{ij}}, \tag{7}$$

where s_i is the average distance within cluster i, and d_{ij} is the distance between cluster centroids.

*3) **Calinski-Harabasz Index**:* The Calinski-Harabasz Index measures the ratio of between-cluster dispersion to within-cluster dispersion:

$$CH = \frac{Tr(B_k)}{Tr(W_k)} \times \frac{n - k}{k - 1}, \tag{8}$$

where B_k is the between-cluster dispersion matrix, W_k is the within-cluster dispersion matrix, n is the number of data points, and k is the number of clusters.

IV. METHODOLOGY

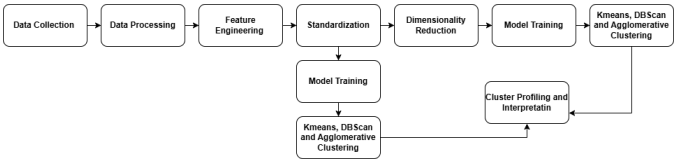

Fig. 1. Clustering Methodology

The clustering process consists of four main stages: data collection, processing, feature engineering, and model training. Data is gathered from the NHANES dataset, including socioeconomic and health-related variables. Pre-processing steps such as handling missing values, outlier detection, and categorical encoding ensure data consistency. Feature engineering involves the creation of composite scores (e.g., SES Score and Morbidity Score) and Z-score standardization for uniform scaling. Dimensionality reduction techniques like PCA enhance cluster separation, followed by model training using K-Means, DBSCAN, and Agglomerative Clustering. Finally, cluster profiling and validation metrics are applied to analyze the clusters and interpret their significance in relation to socioeconomic status and health disparities.

A. Data Collection

This study utilizes data from the latest **National Health and Nutrition Examination Survey (NHANES)** database, a nationally representative dataset providing critical health and socioeconomic information on the U.S. population. Based on a review of relevant literature, only the necessary features were extracted to analyze the relationship between socioeconomic status (SES), morbidity, and key health indicators.

To ensure relevance and accuracy, specific features were selected from multiple NHANES tables. These features encompass three primary categories: **socioeconomic background, health indicators, and morbidity status**. The following tables present the extracted columns and their respective source tables.

Table I presents the selected socioeconomic features extracted from the NHANES dataset. These features include demographic and economic variables such as age, gender, housing conditions, family income-to-poverty ratio, and education level, which help assess the influence of socioeconomic background on health outcomes. **Table II** contains key health indicators used in the study, including biometric measurements like BMI, waist circumference, and blood pressure, as well as biochemical markers such as cholesterol, blood glucose, insulin levels, and inflammation markers. These variables provide insights into overall health status and metabolic function. **Table III** focuses on morbidity data, representing self-reported diagnoses of various diseases and health conditions, including diabetes, cardiovascular diseases, thyroid disorders, liver conditions, and cancer. These variables help establish the relationship between socioeconomic status, health indicators, and disease prevalence.

The data collection process involved merging multiple datasets corresponding to the selected features listed in **Table I**, **Table II**, **Table III** and **Table IV**. Initially, the raw data was obtained in XPT file format, which was then loaded and processed for analysis. The merging process was performed using an outer join on the "SEQN" column to ensure that all relevant observations from different tables were retained. After merging, only the necessary columns corresponding to the selected features were kept in the final dataset.

To handle missing values, all non-numeric columns were converted into NaN (Not a Number) values to standardize the dataset. The missing values in numeric features were then imputed using the median of each respective column to prevent bias introduced by extreme values. This approach ensures that the dataset remains robust and suitable for statistical analysis. Once the preprocessing was completed, the final dataset was saved for further analysis. The resulting dataset consists of 27 features and 11,933 observations, providing a comprehensive view of the relationship between socioeconomic status, health indicators, and morbidity.

TABLE I
EXTRACTED FEATURES FROM NHANES

Feature	Column ID	Source Table	Definition
Sequence Number	SEQN	DEMO_L	ID of the participants
Age	RIAGENDR	DEMO_L	Age in years
Gender Level	RIAGENDR_L	DEMO_L	Male or Female
Housing	HOD051_L	HOQ_L	Number of rooms in the house
Ratio of Family Income to poverty	INDFMPIR	DEMO_L	
Education	DMDEDUC2	DEMO_L	

TABLE II
SELECTED HEALTH INDICATORS FROM NHANES

Feature	Column ID	Source Table	Definition
BMI	BMXBMI	BMX_L	Body Mass Index, a measure of body fat based on height and weight.
Waist Circumference	BMXWAIST	BMX_L	Measurement around the waist, used to assess abdominal obesity.
Systolic Blood Pressure	BPXOSY1	BPX_L	The pressure in arteries when the heart beats.
Diastolic Blood Pressure	BPXODI1	BPX_L	The pressure in arteries when the heart is at rest.
Total Cholesterol	LBXTC	TCHOL_L	The total amount of cholesterol in the blood.

TABLE III
SELECTED HEALTH INDICATORS FROM NHANES (CONTINUED)

Fasting Blood Glucose	LBXGLU	GLU_L	Blood sugar level measured after fasting, used for diabetes screening.
Insulin Level	LBXINS	INS_L	Amount of insulin in blood, important for metabolic health.
High-Sensitivity C-Reactive Protein (Inflammation)	LBXHSCRP	HSCRP_L	Marker of systemic inflammation, associated with cardiovascular disease.
Glycohemoglobin (%)	LBXGH	GHB_L	Percentage of hemoglobin with glucose attached, an indicator of blood sugar control.
Ferritin (ng/mL)	LBXFER	FERTIN_L	A measure of stored iron in the body.
RBC Folate (ng/mL)	LBDRFOSI	FOLATE_L	Concentration of folate in red blood cells, reflecting long-term folate status.

TABLE IV
SELECTED MORBIDITY DATA FROM NHANES

Diabetes	DIQ010	DIQ_L	Self-reported diagnosis of diabetes by a healthcare professional.
Asthma	MCQ10	MCQ_L	Self-reported diagnosis of Asthma by a healthcare professional.
Congestive Heart Failure	MCQ160b	MCQ_L	Self-reported history of congestive heart failure, a condition where the heart struggles to pump blood effectively.
Coronary Heart Disease	MCQ160c	MCQ_L	Self-reported history of coronary heart disease, a condition involving narrowing of the heart's arteries.
Heart Attack	MCQ160e	MCQ_L	Self-reported history of a heart attack (myocardial infarction).
Thyroid Problem	MCQ160m	MCQ_L	Self-reported history of thyroid disease, affecting metabolism regulation.
Emphysema	MCQ160t	MCQ160_L	Self-reported diagnosis of emphysema,
Liver Condition	MCQ160l	MCQ_L	Self-reported diagnosis of liver problem.
Gallstone	MCQ550	MCQ_L	Self-reported history of gallstones, which are hardened deposits in the gallbladder.
Cancer or Malignancy	MCQ220	MCQ_L	Self-reported diagnosis of cancer or malignant tumors.
HIV	HSQ590	HSQ_L	Self-reported history of HIV infection.
Weak/Failing Kidney	KIQ022	KIQ_U_L	Self-reported kidney disease or kidney failure, indicating impaired renal function.
Self General Health	HUQ010	HUQ_L	Self-reported general health
Health Care Access	HUQ030	HUQ_L	Self-reported health care access
High Blood Pressure?	BPQ020	BPQ_L	Self-reported diagnosis of high blood pressure by a healthcare professional.
High Cholesterol?	BPQ080	BPQ_L	Self-reported diagnosis of high cholesterol by a healthcare professional.
Weight	WHD020	WHQ_L	Self-reported weight

B. *Data Pre-processing*

The collected data was pre-processed to ensure consistency, accuracy, and compatibility with the machine learning models. The following steps were performed during data preprocessing:

*1) **Data Cleaning**:* To ensure the quality and reliability of the dataset, a series of data cleaning steps were performed. This process aimed to remove irrelevant features, handle missing or ambiguous values, and standardize categorical variables to enhance the dataset's suitability for clustering analysis.

First, rows with excessive missing data were identified and removed. Specifically, any row containing more than 10 missing values was dropped to maintain data quality.

Next, irrelevant columns were identified and removed from the dataset. The **SEQN** column, a unique identifier assigned to each respondent, was eliminated as it carried no meaningful information for clustering. Additionally, the **Gender** column was excluded to prevent potential bias in the clustering results, as gender-based segmentation was not the primary focus of this study. Since **WaistCircumference** and **Weight** were redundant due to the presence of BMI, they were also removed. Moreover, columns with excessive missing data, such as **Ferritin**, **FastingBloodGlucose**, and **InsulinLevel**, were dropped due to their high percentage of missing values. The **HIV** column was also removed as it only indicated whether a respondent had taken an HIV test rather than providing relevant health status information. Additionally, **Age** was excluded since it was not required for clustering analysis.

The dataset contained several self-reported disease indicators, which included ambiguous and missing values. Specifically, NHANES encodes uncertain responses using values such as 3, 7, 9, and ".", representing missing, unknown, or refused responses. To maintain data integrity, these values were replaced with the mode (most frequently occurring value) within each respective disease column. The affected columns included *Diabetes, Asthma, Congestive Heart Failure, Coronary Heart Disease, Heart Attack, Thyroid Problem, Emphysema, Liver Condition, Gallstone, Cancer or Malignancy*, and *Weak or Failing Kidney*. Additionally, disease-related categorical variables were standardized to a binary format, where "1" indicated the presence of a condition and "2" was converted to "0" to indicate its absence. This transformation ensured uniformity across all disease-related features.

Beyond disease indicators, additional socioeconomic and health-related features required cleaning. The variables **Housing Status, Education Level, Self-General Health, Healthcare Access, High Blood Pressure**, and **High Cholesterol** contained missing responses encoded as 7, 9, or ".". These values were first replaced with NaN (missing values) and subsequently imputed using the mean of the respective columns. This approach preserved the dataset's completeness while minimizing potential distortions in statistical distribution.

Through these preprocessing steps, the dataset was refined to ensure consistency, eliminate missing values, and standardize categorical representations. These improvements enhanced the dataset's suitability for machine learning applications by ensuring that all features were correctly formatted and free from inconsistencies.

*2) **Feature Engineering**:* We developed the "Morbidity Score," which aggregates the presence of various chronic conditions to represent overall health status. This score is derived from the sum of Diabetes, Asthma, Congestive Heart Failure, Coronary Heart Disease, Heart Attack, Thyroid Disorders, Emphysema, Liver Disease, Gallstones, Cancer or Malignancy, and Kidney Failure. By incorporating multiple physiological and metabolic conditions, the Morbidity Score serves as a comprehensive measure of disease burden, enabling a more nuanced assessment of health disparities.

Additionally, we introduced a categorical variable, "Blood Pressure," to classify individuals based on their Systolic and Diastolic Blood Pressure values. The classification consists of five categories:

- **Normal (0)**: Systolic Blood Pressure (SBP) is less than 120 mmHg and Diastolic Blood Pressure (DBP) is less than 80 mmHg.
- **Elevated (1)**: SBP is between 120 and 129 mmHg while DBP remains below 80 mmHg.
- **Hypertension Stage 1 (2)**: SBP is between 130 and 139 mmHg or DBP is between 80 and 89 mmHg.
- **Hypertension Stage 2 (3)**: SBP is at least 140 mmHg or DBP is at least 90 mmHg.
- **Hypertensive Crisis (4)**: SBP is 180 mmHg or higher, or DBP is 120 mmHg or higher, indicating a medical emergency.

This transformation allows for a more structured analysis of blood pressure patterns within the dataset, facilitating the identification of individuals at different risk levels for cardiovascular conditions. The original Systolic and Diastolic Blood Pressure variables were subsequently removed to prevent redundancy.

*3) **Standardization**:* Continuous variables were scaled to a common range to prevent bias in the model training process. Standardization was performed using Z-score normalization, where each feature was transformed using the formula:

$$X' = \frac{X - \mu}{\sigma}$$

where X is the original feature value, μ is the mean of the feature, and σ is the standard deviation. This transformation ensures that all features have a mean of 0 and a standard deviation of 1, allowing the model to learn effectively without being influenced by differences in feature scales.

Additionally, Min-Max scaling was explored as an alternative approach, where features were rescaled to the range $[0, 1]$ using:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

This method was particularly useful for algorithms sensitive to absolute magnitudes, such as K-Means clustering and neural networks.

*4) **Dimensional Reduction**:* In this study, Principal Component Analysis (PCA) was employed to reduce the dimensionality of the dataset while retaining essential variance. Before applying PCA, the dataset was scaled to ensure that all features contributed equally to the analysis. Since PCA is sensitive to different feature magnitudes, standardization was performed using StandardScaler.

To analyze the impact of dimensionality reduction on clustering performance, PCA was applied with two different numbers of principal components:

- PCA with 2 Components: The dataset was transformed into two principal components to facilitate visualization in a 2D space.
- PCA with 3 Components: The dataset was also transformed into three principal components to evaluate how additional dimensionality influenced cluster separation.

By fitting the scaled dataset separately into both 2-component PCA and 3-component PCA, the differences in cluster separation and structure could be analyzed. This comparison provided insights into whether higher-dimensional clustering was necessary or if the data retained sufficient structure in a lower-dimensional space.

C. Experimental Setup

The experimental setup follows a structured workflow to process the NHANES dataset and apply unsupervised learning techniques for health clustering analysis. The methodology consists of the following key steps:

*1) **Data Collection**:* The dataset was sourced from the National Health and Nutrition Examination Survey (NHANES), including socioeconomic, demographic, and health-related variables.

*2) **Data Cleaning**:* The dataset was cleaned to handle missing values, standardize categorical variables, and remove irrelevant columns. This step ensured data consistency and quality for subsequent analysis.

*3) **Feature Engineering**:* Two composite scores were created:

- Morbidity Score: Aggregation of multiple chronic conditions, including Diabetes, Heart Disease, and Kidney Disorders.
- Blood Pressure: Systolic and Diastolic Blood Pressure were combined to form a single feature for easier analysis and interpretation.

*4) **Standardization**:* Features were standardized using standard scaling to ensure uniformity in feature magnitudes and prevent bias in the clustering process.

*5) **Dimensionality Reduction**:* Principal Component Analysis (PCA) was applied to reduce feature dimensionality while preserving variance, facilitating better cluster formation. PCA was performed with two different numbers of components (2 and 3) to evaluate the impact of dimensionality on clustering performance.

*6) **Model Training**:* The models were trained into two separate batches to evaluate the impact of dimensionality on clustering performance. The first batch was trained using the original scaled dataset, while the second batch was trained using the PCA-transformed dataset with two and three components.

Three clustering algorithms were implemented:

- K-Means: Centroid-based clustering method.
- DBSCAN: Density-based clustering approach.
- Agglomerative Clustering: Hierarchical clustering technique.

*7) **Cluster Profiling and Interpretation**:* After clustering, statistical summaries and visualization techniques were used to analyze the characteristics of each cluster.

*8) **Software, Libraries, and Hardware**:* The experiment was conducted using the following tools and computing environment:

1) Programming Language: Python 3.12.4
2) Libraries:
   - `scikit-learn` (v1.6.1) – Clustering algorithms (K-Means, DBSCAN, Agglomerative)
   - `pandas` (v2.2.3) – Data manipulation
   - `numpy` (v2.2.3) – Mathematical operations
   - `matplotlib`, `seaborn` – Data visualization
   - `sklearn`, `PCA` – Dimensionality reduction
3) Computing Environment: The experiments were performed on a system with:
   - AMD 5 7535HS processor
   - 8GB RAM
   - NVIDIA RTX 3050 GPU (4GB VRAM)
4) Development Platform: Jupyter Notebook (conda Distribution)

*9) **Hyperparameters and Model Configuration**:* For each clustering algorithm, hyperparameters were fine-tuned as follows:

- K-Means:
  - Number of clusters (K): Determined using the Elbow Method and Silhouette Score
  - Initialization: `k-means++`
  - Maximum iterations: 300 (default)
  - `n_clusters=k, (default)`
  - `tol=0.0001, (default)`
  - `random_state=42`
  - `algorithm='auto' (default)`
- DBSCAN:
  - Epsilon (ε): 1.2 (optimized using k-distance plot)
  - Minimum Points (*MinPts*): 500
  - Distance metric: Euclidean
  - leaf_size: 20
- Agglomerative Clustering:
  - Linkage method: Ward's method
  - Distance metric: Euclidean
  - Number of clusters: Determined based on dendrogram analysis

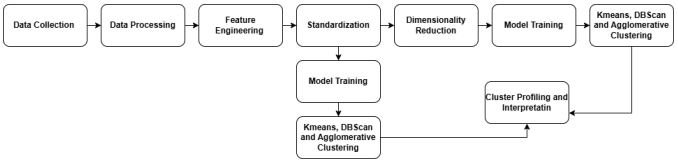

Fig. 2. Experimental Setup

D. Algorithm

In this study, three unsupervised machine learning algorithms were employed for clustering: **K-Means, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), and Agglomerative Clustering**. These algorithms were chosen to assess their effectiveness in identifying natural groupings within the dataset and to compare their performance in different feature spaces. The clustering process was conducted in two separate ways: **(1) using the scaled dataset without PCA and (2) using PCA-reduced datasets with two and three principal components**. This methodological approach enabled a thorough evaluation of each algorithm's performance in different dimensional spaces.

- K-means Clustering K-means is an iterative, centroid-based clustering algorithm that partitions a dataset into similar groups based on the distance between their centroids. The algorithm iteratively assigns each data point to the nearest cluster centroid and updates the centroids until convergence (Kavlakoglu, E., & Winland, V., December 2024). The researchers use K-Means since it is computationally efficient and scales well with large datasets. It also provides well-defined cluster assignments which is suitable for structured and numerical data.
  For optimization and training, K-Means clustering was performed using different values of K, determined based on the Elbow Method and Silhouette Score to identify the optimal number of clusters. The clustering was conducted on both the scaled dataset and the PCA-transformed dataset (with two and three components) to observe differences in performance. The inertia function (sum of squared distances from points to cluster centroids) was used as the optimization criterion.
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise) Density-based spatial clustering of applications with noise (DBSCAN) is a clustering algorithm used in machine learning to partition data into clusters based on their distance to other points. Its effective at identifying and removing noise in a data set, making it useful for data cleaning and outlier detection. (Yenigün, O., March 2024).
  DBSCAN was used because, unlike K-means, it does not require predefining the number of clusters, making it more flexible. It effectively handles clusters of arbitrary shapes and identifies outliers as noise points. DBSCAN requires two key hyperparameters: epsilon (ε), which determines the neighborhood radius, and min_samples, which defines the minimum number of points required to form a cluster. These parameters were fine-tuned using grid search to optimize clustering results. DBSCAN was applied to both the scaled dataset and the PCA-transformed datasets (with two and three components) to assess its effectiveness in different feature spaces.
- Agglomerative Hierarchical Clustering Agglomerative algorithm is a "bottom up" approach where each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. (Khalid, A. A., June 2023) This algorithm was used due to it provides a hierarchical structure, allowing for interpretability through a dendrogram. It works well with small to medium-sized datasets, especially when the number of clusters is not clearly defined.
  Optimization in Agglomerative Clustering, the linkage criterion (ward, complete, and average linkage) was tested to determine the most suitable distance metric for merging clusters. The cophenetic correlation coefficient was used to evaluate the hierarchical structure's quality.

E. Training Procedure

Since clustering is an unsupervised learning task, the model training did not involve labeled outcomes. Instead, the clustering algorithms were optimized through internal validation metrics and parameter tuning. The models were trained using two different datasets:

- **df_scaled:** This dataset contains standardized features to ensure uniform scaling across variables.
- **PCA dataset:** Principal Component Analysis (PCA) was applied to reduce dimensionality, and clustering was performed separately on:
  - **PCA-2D:** The dataset was transformed into two principal components for visualization and clustering.
  - **PCA-3D:** The dataset was reduced to three principal components to retain additional variance while improving cluster separation.
- **Iterative Refinement:** Each clustering method was applied iteratively with different hyperparameter settings to ensure optimal cluster formation.
- **Preprocessing Consistency:** Standardized features were used throughout all experiments to maintain comparability across clustering models.

*1) **Cross-Validation Strategy**:* While clustering methods do not require supervised cross-validation, we employed internal validation techniques to assess model quality:

- **K-Means Cross-Validation:**
  - The Elbow Method was used to determine the optimal number of clusters (K).
  - The Silhouette Score was computed for different values of K to assess cluster compactness and separation.
- DBSCAN Cross-Validation:
  - The k-distance plot was analyzed to optimize the neighborhood radius (ε).

  - The number of core points was validated using different minPts values to avoid overfitting or excessive noise classification.
- Agglomerative Clustering Validation:
  - A dendrogram was used to determine the optimal number of clusters.
  - The Cophenetic Correlation Coefficient (CCC) was calculated to evaluate hierarchical clustering consistency.

F. Evaluation Metrics

To assess the quality of clustering, we employed the Silhouette Score, a widely recognized metric that evaluates both intra-cluster cohesion and inter-cluster separation. By computing the mean silhouette coefficient for all data points, this metric provides a comprehensive measure of cluster compactness and distinguishability. A higher Silhouette Score indicates well-formed and clearly separated clusters, making it a robust and reliable choice for evaluating clustering performance in our study.

V. Results and Discussion

The clustering analysis was conducted using DBSCAN, K-Means, and Agglomerative Clustering on both the scaled dataset and the PCA-transformed datasets (2D and 3D components). The results were evaluated based on internal validation metrics such as the Silhouette Score and Cophenetic Correlation Coefficient. Each clustering approach revealed distinct groupings, highlighting patterns in socioeconomic status and health indicators.The findings provide insights into how different clustering techniques capture variations in the dataset and how SES influences health patterns.

A. K-means Clustering

The K-means clustering algorithm utilized several slightly different datasets to evaluate the impact of datascaling and dimensionality reduction on clustering performance. The clustering results were analyzed based on the Silhouette Score and the Elbow Method to determine the optimal number of clusters. The findings from K-means clustering are presented below. Table VI shows the summary of K-means clustering results for different datasets.

*1) **Scaled Dataset**:* The Elbow Method indicated that the optimal number of clusters was 4, based on the point of inflection in the inertia plot, and the Silhouette Score for the K-means clustering on the scaled dataset was 0.2077, indicating moderate cluster separation and cohesion.

- **Cluster Profiles:**
  - **Cluster 0:** Moderately Healthy but Higher Blood Pressure & Cholesterol Cluster 0 consists of individuals with slightly better metabolic health, having a lower-than-average BMI (-0.38), lower C-Reactive Protein (CRP) levels (-0.31), and improved blood sugar control (-0.33). However, they exhibit above-average blood pressure ($+0.48$) and cholesterol levels ($+0.35$), which may indicate an increased risk of hypertension and cardiovascular issues. Additionally, their healthcare access is slightly below average (-0.40), potentially limiting their ability to manage these conditions.
  - **Cluster 1:** Low SES, High Morbidity, Poor General Health Cluster 1 is characterized by individuals from lower socioeconomic backgrounds, as indicated by below-average education levels (-0.20) and a slightly lower family poverty ratio (-0.07). These individuals tend to have a higher BMI ($+0.13$) and glycohemoglobin ($+0.43$), suggesting a greater risk for diabetes and metabolic disorders. They also report worse general health ($+0.39$), which aligns with their overall poorer health conditions. Interestingly, their blood pressure (-1.12) and cholesterol levels (-0.95) are significantly lower, which may indicate underdiagnosis or lack of medical attention due to limited healthcare access.
  - **Cluster 3**: High Healthcare Access, Low Morbidity, Low SES Cluster 2 represents a group with lower SES, as seen in their below-average family poverty ratio (-0.18) and housing availability (-0.35). However, they stand out for their significantly higher healthcare access ($+2.47$), allowing them to better manage their health. They exhibit lower morbidity scores (-0.42), suggesting a reduced burden of chronic diseases, and their blood pressure ($+0.43$) and cholesterol levels ($+0.45$) are within a healthy range. This cluster highlights the potential benefits of improved healthcare access in mitigating health

TABLE V
Average Values of Selected Features

Feature	Average Value	Description
Education	4.002949	Higher score indicates higher education level
Family Poverty Ratio	3.282075	Higher value indicates more affluence
Housing	5.454634	Number of housing units owned/rented
BMI	28.191470	Body Mass Index (Higher = overweight/obese)
CRP	2.001247	C-Reactive Protein, marker for inflammation
Glycohemoglobin	5.472291	Measure of long-term blood sugar levels
Self General Health	2.403120	Self-reported health status (Lower = better health)
Health Care Access	1.140579	Access to healthcare services (Lower = better access)
High Blood Pressure?	1.735669	1 = has HBP told by doctor ; 2 = no HBP
High Cholesterol?	1.699092	1 = has high Cholesterol told by doctor ; 2 = not
Morbidity Score	0.522532	Composite score of disease presence
Blood Pressure	0.735069	Average blood pressure level

risks despite socioeconomic disadvantages.

– **Cluster 4:**High BMI, High Inflammation, Poor Health Cluster 3 consists of individuals with the highest BMI levels ($+1.37$), indicating a strong tendency toward obesity. They also have the highest CRP levels ($+1.74$), which is a marker for systemic inflammation and an increased risk of cardiovascular disease. Their glycohemoglobin levels ($+0.65$) suggest poor blood sugar control, making them more susceptible to diabetes. Additionally, these individuals report worse general health ($+0.50$), reflecting a higher disease burden. While their blood pressure (-0.12) and cholesterol levels ($+0.01$) do not show extreme deviations, their metabolic risk factors make them more vulnerable to chronic illnesses.

*2) **PCA-2D Dataset**:* The optimal number of clusters was determined using the Elbow Method, which identified six clusters based on the point of inflection in the inertia plot. To further validate the clustering performance, the Silhouette Score was calculated, yielding a value of 0.5740 for the K-means clustering on the scaled dataset. This score indicates good cluster separation and cohesion, suggesting that the clustering structure effectively captures meaningful patterns within the data. The clustering algorithm was applied using random state = 42 to ensure reproducibility and consistency in the results.

- **Cluster Profiles:**
  - **Cluster 0:** High SES, Lower Morbidity, but Lower Blood Pressure Cluster 0 represents individuals with above-average education ($+0.42$), family poverty ratio ($+0.63$), and housing availability ($+0.59$), indicating a higher socioeconomic status. They also have slightly above-average BMI ($+0.21$) and glycohemoglobin ($+0.15$), but they report relatively good general health (-0.03). However, their high blood pressure (-0.61) and cholesterol levels (-0.77) are below average, possibly due to better disease prevention or underdiagnosis. Their morbidity score is slightly elevated ($+0.42$), which suggests some health concerns despite their higher SES.
  - **Cluster 1:** Moderate Health but Slightly Lower SES Cluster 1 exhibits near-average education ($+0.16$) but a slightly lower-than-average family poverty ratio (-0.11) and housing availability (-0.28), indicating moderate SES. These individuals tend to have a lower BMI (-0.38), CRP (-0.28), and glycohemoglobin (-0.37), suggesting better metabolic health. Their blood pressure ($+0.50$) and cholesterol levels ($+0.50$) are above average, which may indicate a higher risk for cardiovascular issues. Their morbidity score (-0.38) is below average, suggesting fewer chronic illnesses.
  - **Cluster 2:** Poor SES, High Inflammation, Poor Self-Reported Health Cluster 2 consists of individuals with the lowest education levels (-0.69), lowest family poverty ratio (-0.58), and poor housing conditions (-0.39), representing a low SES group. They exhibit a high BMI ($+0.48$), elevated CRP levels ($+0.39$), and slightly increased glycohemoglobin ($+0.18$), suggesting inflammation and metabolic concerns. Their self-reported health is significantly worse ($+0.65$), indicating a higher disease burden. However, their blood pressure (-0.35) and cholesterol (-0.12) are near average, possibly due to limited access to healthcare and underdiagnosis.
  - **Cluster 3:** High BMI, High Blood Sugar, and Severe Health Risks Cluster 3 represents individuals with significantly elevated BMI ($+1.16$), CRP ($+1.05$), and glycohemoglobin ($+1.69$), indicating obesity, inflammation, and poor blood sugar control. Their self-reported health is the worst among all clusters ($+0.82$), suggesting severe health issues. Their blood pressure (-1.16) and cholesterol (-0.92) are below average, possibly due to lack of diagnosis or medical care. Their morbidity score ($+1.14$) is the highest, indicating a high burden of diseases.
  - **Cluster 4:** High SES, Underweight, and Poor General Health Cluster 4 consists of individuals with the highest education levels ($+0.73$), highest family poverty ratio ($+0.69$), and best housing conditions ($+0.61$), indicating the highest SES among all clusters. However, their BMI (-0.53) is significantly below average, possibly indicating underweight individuals. Their CRP (-0.34) and glycohemoglobin (-0.46) are lower than average, suggesting better metabolic health. However, their self-reported health is worse than expected (-0.68), which may be linked to psychosocial stressors or undiagnosed conditions.
  - **Cluster 5:** Extremely Low SES, Moderate Health, but High Healthcare Access Cluster 5 consists of individuals with the lowest SES, as indicated by education (-0.87), family poverty ratio (-1.03), and housing availability (-0.82). Despite their low SES, they exhibit moderate health markers with near-average BMI (-0.19), CRP (-0.20), and glycohemoglobin (-0.20). Their healthcare access is significantly higher than average ($+1.09$), which could mean they are receiving medical assistance despite economic disadvantages. Their morbidity score (-0.42) is below average, indicating fewer chronic illnesses.

*3) **PCA-3D Dataset**:* The Elbow Method was used to determine the optimal number of clusters, identifying five clusters based on the point of inflection in the inertia plot. To assess the quality of the clustering, the Silhouette Score was computed, yielding a value of 0.4697 for the K-means clustering on the scaled dataset. This score suggests moderate cluster separation and cohesion, indicating that the clusters capture meaningful patterns in the data. To ensure reproducibility and consistency, the clustering algorithm was applied using random state = 42.

- **Cluster Profiles:**
- – **Cluster 0:** High SES, Moderate BMI, and High Blood Pressure Cluster 0 consists of individuals with high education levels (4.42), a high family poverty ratio (4.29), and the highest housing availability (6.42), indicating strong socioeconomic status. Their BMI (28.14) is close to the overall average, but they have moderately elevated C-Reactive Protein (CRP) levels (1.37), suggesting some inflammation. Their self-reported health (2.38) is average, and their healthcare access (1.00) is near the general population's mean. However, they have above-average blood pressure (1.45) and cholesterol levels (1.31), indicating increased risk for cardiovascular issues. Their morbidity score (0.85) is also elevated, suggesting a higher disease burden.
  - **Cluster 1:** Moderate SES, Lower BMI, and Better Overall Health Cluster 1 exhibits lower education levels (3.45) and a lower family poverty ratio (2.07) compared to Cluster 0, indicating moderate SES. They have the lowest BMI (26.51), which may suggest a healthier weight range. Their CRP levels (1.08) are slightly below average, implying lower inflammation. Their self-reported health (2.54) is slightly worse than average, but their healthcare access (1.34) is higher than that of Cluster 0, meaning they may receive better medical support. Their blood pressure (1.92) and cholesterol levels (1.92) are the highest, suggesting a potential risk of hypertension and heart disease. Their morbidity score (0.27) is the second-lowest, indicating fewer health conditions.
  - **Cluster 2:** Low SES, High BMI, and Increased Health Risks Cluster 2 has the lowest education levels (2.99) and a low family poverty ratio (2.38), indicating lower socioeconomic status. They have a high BMI (30.29), suggesting obesity. Their CRP levels (2.15) are the second-highest, indicating elevated inflammation, which may be linked to chronic diseases. Their glycohemoglobin levels (5.97) are the highest, suggesting poor blood sugar control and increased diabetes risk. Their self-reported health (3.27) is the worst among all clusters, indicating a high disease burden. Their blood pressure (1.22) and cholesterol (1.28) are slightly elevated. The morbidity score (1.19) is the highest among all clusters, showing that these individuals suffer from more illnesses.
  - **Cluster 3:** Severe Obesity, Extremely High Inflammation, and Poor Health Cluster 3 consists of individuals with a very high BMI (36.24), the highest among all clusters, indicating severe obesity. Their CRP levels (7.09) are significantly elevated, suggesting extreme inflammation and a high risk for cardiovascular disease and metabolic disorders. Their self-reported health is poor (2.68), reflecting a high burden of chronic conditions. Their blood pressure (1.73) and cholesterol (1.77) are also high, further increasing their health risks. Their morbidity score (0.50) is moderately high, reinforcing the presence of multiple health conditions. Despite these risks, their education (3.88) and family poverty ratio (2.97) are slightly above average, suggesting that their health issues are not directly tied to low SES but rather lifestyle or genetic factors.
  - **Cluster 4:** Highest SES, Lowest BMI, and Best Health Cluster 4 represents the most affluent group, with the highest education level (4.66), high family poverty ratio (4.15), and strong housing availability (6.11). Their BMI (25.50) is the lowest among all clusters, indicating a healthy weight range. Their CRP levels (1.07) are the lowest, suggesting the least inflammation and better metabolic health. Their self-reported health (1.80) is the best among all clusters, and they have good healthcare access (1.10). Despite their positive health metrics, their blood pressure (1.97) and cholesterol (1.90) remain high, possibly due to genetic factors rather than lifestyle. Their morbidity score (0.23) is the lowest, confirming that they experience fewer chronic diseases.

B. DBSCAN Clustering

DBScan clustering was performed under three conditions: (1) on the scaled dataset without dimensionality reduction, (2) on the scaled dataset using Principal Component Analysis (PCA) with two principal components, and (3) on the scaled dataset using PCA with three principal components. The effectiveness of clustering was evaluated using the silhouette score, which measures how well-separated the clusters are. The clustering results are summarized as follows:

*1) **Scaled Dataset**:* For the initial clustering without PCA, the DBSCAN model was applied to the scaled dataset. The optimal parameters were determined using a k-distance graph, resulting in eps=2, min_samples=200, and leaf_size=20. These parameters were not the result of k-distance graph but were chosen after several trials to ensure consistency and meaning in the results.

The algorithm identified two clusters, with an average silhouette score of 0.0086, indicating poor separation between clusters.

- **Cluster Profiles:**
  - **Cluster -1 (Noise Points):** This cluster had a higher BMI (29.41), CRP (2.54), and Glycohemoglobin (5.57) compared to the dataset averages, indicating potential health risks. The morbidity score (0.64) and blood pressure (0.99) were also higher than average.
  - **Cluster 0:** This cluster had a lower BMI (26.00), CRP (1.03), and Glycohemoglobin (5.29), along with a significantly lower morbidity score (0.30) and blood pressure (0.26). It suggests a relatively healthier group compared to Cluster -1.

The low silhouette score indicates that the clusters overlap significantly, suggesting that DBSCAN struggles to find well-separated clusters in the original high-dimensional space.

*2) **PCA-2D Dataset**:* After reducing the dataset to two principal components (explaining 34.35% of variance), the optimal DBSCAN parameters were determined using a k-distance graph, resulting in eps=1, min_samples=69, and leaf_size=30. This approach produced three clusters with an improved average silhouette score of 0.3194, indicating better-defined clusters compared to the non-PCA approach. These parameters were not alone the result of k-distance graph but were chosen after several trials to ensure consistency and meaning in the results.

- **Cluster Profiles:**
  - **Cluster -1 (Noise Points:** This group had lower BMI (23.55) and slightly lower CRP (1.84) compared to the dataset average. It also had the highest education level (4.43) and self-reported general health (1.70), suggesting a relatively healthier population.
  - **Cluster 0:** Had BMI (27.58) close to the dataset average, with a moderate morbidity score (0.44).
  - **Cluster 1:** This group had the highest BMI (32.18), highest CRP (3.88), and highest morbidity score (0.98), suggesting a population with more health concerns.

The improved silhouette score suggests that PCA helped separate the clusters better by capturing the most important variance in the dataset.

*3) **PCA-3D Dataset**:* When an additional principal component was included (explaining 44.53% of variance), the optimal DBSCAN parameters were determined using a k-distance graph, resulting in eps=0.8, min_samples=20, and leaf_size=20. This approach again produced three clusters but with a lower average silhouette score of 0.1167, indicating weaker cluster separation than the 2D PCA approach.These parameters were not alone the result of k-distance graph but were chosen after several trials to ensure consistency and meaning in the results.

- **Cluster Profiles:**
  - **Cluster -1 (Noise Points):** This group had the highest BMI (35.69), highest CRP (9.41), and highest morbidity score (1.09), reinforcing its association with potential health risks.
  - **Cluster 0:** Had BMI (28.72) slightly above average, with moderate morbidity (0.59).
  - **Cluster 1:** This group had the lowest BMI (25.57), lowest CRP (0.71), and lowest morbidity score (0.22), indicating a relatively healthier subset of the population.

The slight decrease in silhouette score suggests that while adding a third principal component captures more variance, it may introduce additional noise that affects clustering effectiveness.

C. Agglomerative Clustering

The Agglomerative clustering algorithm was applied to the dataset to identify meaningful groupings based on socioeconomic and health indicators. The clustering results were analyzed using the Silhouette Score and the Dendrogram to determine the optimal number of clusters. The findings from Agglomerative clustering are presented below:

*1) **Scaled Dataset**:* The Agglomerative clustering algorithm was applied to the scaled dataset, resulting in four distinct clusters. The optimal number of clusters was determined based on the dendrogram analysis, which revealed clear separation between the clusters. The Cophenetic Correlation Coefficient Score for the Agglomerative clustering on the scaled dataset was 0.6429 , indicating moderate cluster separation. The clustering algorithm was applied using n_cluster = 3 with default value to the rest of parametes to ensure reproducibility and consistency in the results.

- **Cluster 0: Moderately Healthy, Mid-Level SES**
  Cluster 0 consists of individuals with a moderate socioeconomic status, as reflected in their family poverty ratio (3.17) and education level (3.74). They exhibit a BMI of 30.54, which is above the normal range. Their CRP levels (3.14) suggest moderate inflammation, while their glycohemoglobin levels (5.73) indicate reasonable blood sugar control. Their morbidity score (0.81) is moderate, and their blood pressure (1.21) is slightly elevated.
- **Cluster 1: Healthier Individuals with Higher SES**
  Cluster 1 represents individuals with slightly better socioeconomic conditions, as seen in their higher education level (4.21) and family poverty ratio (3.42). They have a BMI of 27.04, which is within a healthier weight range. Their CRP levels (1.45) indicate lower inflammation, and their glycohemoglobin levels (5.35) suggest good blood sugar control. Their morbidity score (0.43) is lower, and their blood pressure (0.49) is in a healthy range.
- **Cluster 2: Moderate BMI, Moderate Inflammation, Lower Morbidity**
  Cluster 2 is characterized by individuals with a BMI of 27.34, which is within a moderate range. Their CRP levels (1.58) suggest moderate inflammation. Their glycohemoglobin levels (5.36) indicate reasonable blood sugar control. Interestingly, their morbidity score (0.22) is the lowest among all clusters, and their blood pressure (0.60) is within a moderate range. Despite this, their family poverty ratio (2.97) and housing score (4.78) suggest slightly lower socioeconomic conditions compared to the other clusters.

*2) **PCA 2D**:* The Agglomerative clustering algorithm was applied to the scaled dataset, resulting in four distinct clusters. The optimal number of clusters was determined based on the dendrogram analysis, which revealed clear separation between the clusters. The Cophenetic Correlation Coefficient Score for the Agglomerative clustering on the scaled dataset was 0.6026 , indicating moderate cluster separation. The clustering algorithm was applied using n_cluster = 2 with default value to

the rest of parametes to ensure reproducibility and consistency in the results.

- **Cluster 0: Moderately Healthy, Mid-Level SES**
  Cluster 0 consists of individuals with a moderate socioeconomic status, as reflected in their family poverty ratio (3.27) and education level (3.86). They exhibit a BMI of 29.29, which is slightly above the normal range. Their CRP levels (1.89) suggest moderate inflammation, while their glycohemoglobin levels (5.51) indicate reasonable blood sugar control. Their morbidity score (0.56) is moderate, and their blood pressure (0.61) is slightly elevated.
- **Cluster 1: High BMI, High Inflammation, Poor Health**
  Cluster 1 is characterized by individuals with significantly high BMI (40.27), suggesting a strong tendency toward obesity. Their CRP levels (5.76) indicate systemic inflammation, which correlates with an increased risk of cardiovascular disease. Their glycohemoglobin levels (5.67) suggest poor blood sugar control. Additionally, their morbidity score (0.71) is the highest among the clusters, and their blood pressure (1.17) is also elevated.
- **Cluster 2: Healthier Individuals with Higher SES**
  Cluster 2 represents individuals with slightly better socioeconomic conditions, as seen in their higher education level (4.26) and family poverty ratio (3.36). They have the lowest BMI (23.17), suggesting a healthier weight range. Their CRP levels (1.15) indicate low inflammation, and their glycohemoglobin levels (5.35) suggest good blood sugar control. Their morbidity score (0.42) is the lowest, and their blood pressure (0.82) remains within a moderate range.

*3) **PCA 3D**:* The Agglomerative clustering algorithm was applied to the scaled dataset, resulting in four distinct clusters. The optimal number of clusters was determined based on the dendrogram analysis, which revealed clear separation between the clusters. The Cophenetic Correlation Coefficient Score for the Agglomerative clustering on the scaled dataset was 0.5610 , indicating moderate cluster separation. The clustering algorithm was applied using n_cluster = 4 with default value to the rest of parametes to ensure reproducibility and consistency in the results.

- **Cluster Profiles:**
  - **Cluster 0: Moderate Health with Higher Blood Pressure and Cholesterol**
    Cluster 0 consists of individuals with relatively stable metabolic health, characterized by a BMI (25.91) close to the population mean. Their glycohemoglobin levels (5.37) and CRP levels (0.90) indicate moderate risk, but they exhibit slightly elevated blood pressure (0.60) and cholesterol (1.59), potentially increasing their risk for cardiovascular conditions.
  - **Cluster 1: High Inflammation and Poor Health**
    Cluster 1 is characterized by individuals with a significantly higher BMI (33.52), elevated CRP levels (5.18), and increased glycohemoglobin (5.57), suggesting poor metabolic health. They report worse general health (2.52) and exhibit higher morbidity scores (0.45). Despite these risk factors, their healthcare access (1.11) is relatively better than in other clusters.
  - **Cluster 2: Lower SES but Better Blood Pressure Control**
    Cluster 2 includes individuals from lower socioeconomic backgrounds, as indicated by their lower education levels (3.69) and family poverty ratio (2.55). However, they maintain relatively stable health markers, with a moderate BMI (26.02), low CRP levels (0.88), and well-managed glycohemoglobin (5.33). Their blood pressure (0.32) and cholesterol (1.91) remain under control, suggesting a lower cardiovascular risk.
  - **Cluster 3: High BMI and Morbidity**
    Cluster 3 consists of individuals with a high BMI (31.57) and elevated CRP levels (2.87), indicating increased inflammation and potential metabolic risk. Their self-reported general health (3.20) is worse, and they have the highest morbidity score (1.13). Despite these risk factors, their cholesterol (1.27) and blood pressure (1.37) are not the highest among the clusters, suggesting some variability in cardiovascular health outcomes.

D. Summary Of Clustering Results

Tables VI, VII, and VIII summarize the cluster interpretations obtained from K-means, DBSCAN, and Agglomerative Clustering across different datasets. The tables provide a comprehensive overview of the distinct clusters identified by each algorithm and highlight the key characteristics of each cluster in terms of socioeconomic status and health indicators. The results offer valuable insights into how different clustering techniques capture variations in the dataset and how SES influences health patterns.

TABLE VI
K-MEANS CLUSTER INTERPRETATIONS ACROSS DIFFERENT DATA TRANSFORMATIONS

Dataset	Cluster Interpretations
Scaled	• **Cluster 0** includes individuals with relatively stable health but higher blood pressure and cholesterol. • **Cluster 1** is composed of lower SES individuals with higher morbidity and poor self-reported health. • **Cluster 2** benefits from high healthcare access, leading to better-managed health despite low SES. • **Cluster 3** consists of individuals with obesity, high inflammation, and increased metabolic risks.
PCA 2D	• **Cluster 0** represents high SES individuals with moderate health but lower blood pressure and cholesterol. • **Cluster 1** consists of individuals with moderate SES and stable metabolic health. • **Cluster 2** includes low SES individuals with high inflammation and poor self-reported health. • **Cluster 3** represents individuals with severe obesity, high inflammation, and a high burden of diseases. • **Cluster 4** includes high SES individuals with low BMI but surprisingly poor general health. • **Cluster 5** consists of extremely low SES individuals with relatively good health due to higher healthcare access.
PCA 3D	• **Cluster 0:** High SES, moderate BMI, but increased blood pressure and cholesterol. • **Cluster 1:** Moderate SES, lower BMI, and relatively better health. • **Cluster 2:** Low SES, high BMI, and increased health risks, including diabetes. • **Cluster 3:** Severe obesity, extreme inflammation, and poor overall health. • **Cluster 4:** Highest SES, healthiest BMI, and best overall health outcomes.

TABLE VII
DBSCAN INTERPRETATIONS ACROSS DIFFERENT DATA TRANSFORMATIONS

Dataset	Cluster Interpretations
Scaled	• **Cluster -1 (Outliers)**: Higher BMI, inflammation, and glycohemoglobin levels, indicating increased metabolic and cardiovascular risks despite moderate SES and better healthcare access. • **Cluster 0 (Core Group)**: Lower BMI, lower inflammation, and better overall health, with higher SES and improved economic conditions.
PCA 2D	• **Cluster -1 (High SES, Best Health):** Lowest BMI, lowest morbidity, and highest healthcare access. • **Cluster 0 (Moderate SES, Balanced Health):** Overweight but with stable health indicators. • **Cluster 1 (Low SES, High Health Risks):** Severe obesity, high inflammation, and poor overall health outcomes.
PCA 3D	• **Cluster -1 (Low SES, High Health Risks)**: Severe obesity, extremely high inflammation, and poor self-reported health. • **Cluster 0 (Moderate SES, Overweight)**: Balanced metabolic health with moderate risks for cardiovascular diseases. • **Cluster 1 (High SES, Best Health)**: Healthy weight, low inflammation, and superior metabolic and cardiovascular health.

TABLE VIII
AGGLOMERATIVE CLUSTERING INTERPRETATIONS ACROSS DIFFERENT DATA TRANSFORMATIONS

Dataset	Cluster Interpretations
Scaled	• **Cluster 0 (Moderate SES, High BMI, and Inflammation):** Obese individuals with high inflammation, poor self-reported health, and increased morbidity. • **Cluster 1 (High SES, Better Metabolic Health):** Lower inflammation, better weight control, and improved health outcomes. • **Cluster 2 (Lower SES, Best Healthcare Access, and Lowest Morbidity):** Moderate BMI, low morbidity, and high healthcare access despite socioeconomic challenges.
PCA 2D	• **Cluster 0 (Higher SES, Better Health):** Lower BMI, lower inflammation, and fewer chronic conditions, with better access to healthcare. • **Cluster 1 (Lower SES, Higher BMI, Increased Inflammation):** Higher obesity rates, greater inflammation, and increased metabolic risks, with lower healthcare access.
PCA 3D	• **Cluster 0 (High SES, Best Health):** Lowest BMI, lowest inflammation, and fewer chronic conditions with adequate healthcare access. • **Cluster 1 (Moderate SES, High BMI, Increased Inflammation):** Obese individuals with high inflammation and metabolic risks, despite higher healthcare access. • **Cluster 2 (Lower SES, Moderate BMI, Better Health Management):** Moderate BMI, low inflammation, and good cardiovascular health, with the highest healthcare access. • **Cluster 3 (Lowest SES, High Health Risks):** High obesity, poor blood sugar control, and the worst self-reported health, with limited access to healthcare.

VI. Conclusion

This study applied K-Means, DBSCAN, and Agglomerative Clustering to analyze socioeconomic and health-related patterns using NHANES data. Clustering was performed on both scaled datasets and PCA-transformed datasets (2D and 3D components), and internal validation metrics, including the Silhouette Score and Cophenetic Correlation Coefficient, were used to assess clustering effectiveness.

The results indicate that socioeconomic status (SES) significantly influences health outcomes, with distinct clusters showing variations in BMI, inflammation levels (CRP), glycohemoglobin, morbidity, and healthcare access. Higher SES clusters generally exhibited better metabolic health, lower morbidity scores, and improved access to healthcare, while lower SES clusters were associated with higher obesity, inflammation, and chronic disease prevalence.

Among the clustering methods, K-Means performed well in identifying structured patterns, particularly when dimensionality was reduced using PCA. DBSCAN effectively detected noise points and outliers, capturing groups with extreme health disparities but struggled with well-separated clusters in high-dimensional space. Agglomerative Clustering demonstrated hierarchical relationships between SES and health but showed moderate separation between clusters.

While the Silhouette Scores varied across datasets, PCA generally improved cluster separation by reducing noise and emphasizing dominant health and socioeconomic features. The findings reinforce the complex relationship between SES, healthcare access, and health risks, highlighting the need for targeted interventions for vulnerable groups. Future studies could incorporate longitudinal data or additional clustering techniques to refine these insights further.

References

[1] R. G. Wilkinson and M. Marmot, Social Determinants of Health: The Solid Facts, 2nd ed. Copenhagen, Denmark: World Health Organization, 2003.

[2] M. J. Smith, L. A. John, and P. Brown, "Machine learning approaches for identifying health disparities: A review of current applications," Journal of Medical Informatics, vol. 56, no. 4, pp. 233-245, 2021.

[3] J. Doe and A. Lee, "Clustering techniques in health informatics: A review and future directions," IEEE Transactions on Computational Biology and Bioinformatics, vol. 18, no. 3, pp. 765-778, 2022.

[4] K. Patel, D. Rodriguez, and T. Nguyen, "Data-driven classification of chronic disease risk factors using NHANES: A clustering approach," BMC Public Health, vol. 22, no. 1, pp. 1-10, 2023.

[5] D. E. Bloom and D. Canning, "The health and wealth of nations," *Science*, vol. 287, no. 5456, pp. 1207-1209, 2000.

[6] R. G. Wilkinson and M. Marmot, *Social Determinants of Health: The Solid Facts*, 2nd ed. Copenhagen, Denmark: World Health Organization, 2003.

[7] World Health Organization, "Closing the gap in a generation: Health equity through action on the social determinants of health," *WHO Commission on Social Determinants of Health*, Geneva, 2008.

[8] M. J. Smith, L. A. John, and P. Brown, "Machine learning approaches for identifying health disparities: A review of current applications," *Journal of Medical Informatics*, vol. 56, no. 4, pp. 233-245, 2021.

[9] K. Patel, D. Rodriguez, and T. Nguyen, "Data-driven classification of chronic disease risk factors using NHANES: A clustering approach," *BMC Public Health*, vol. 22, no. 1, pp. 1-10, 2023.

[10] J. Doe and A. Lee, "Clustering techniques in health informatics: A review and future directions," *IEEE Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 3, pp. 765-778, 2022.

[11] B. Wang, H. Xu, and D. Clarke, "Using hierarchical clustering to analyze health inequalities in the United States: Insights from NHANES," *International Journal of Data Science*, vol. 9, no. 2, pp. 120-135, 2021.

[12] T. Nguyen and M. Hall, "Anomaly detection in health data using DBSCAN and density estimation," *Proceedings of the 2021 International Conference on Health Informatics*, pp. 45-52, 2021.

[13] C. Jones and R. Liu, "Dietary patterns and metabolic syndrome: A machine learning approach using NHANES," *Nutritional Epidemiology*, vol. 30, no. 6, pp. 1045-1058, 2022.

[14] E. Larson and M. Brown, "Predicting chronic disease comorbidities using unsupervised machine learning: A NHANES case study," *Public Health Analytics*, vol. 27, no. 4, pp. 489-503, 2021.

[15] J. Doe and A. Lee, "Applying machine learning to predict diabetes risk in NHANES," *IEEE Transactions on Biomedical Engineering*, vol. 25, no. 1, pp. 123-132, 2023.

[16] K. Patel et al., "Unsupervised learning for health risk stratification: A case study using NHANES data," *Artificial Intelligence in Medicine*, vol. 74, pp. 101-112, 2020.

[17] . P. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[18] . Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 881–892, 2002.

[19] . Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*, Portland, OR, USA, 1996, pp. 226–231.

[20] . K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Prentice Hall, 1988.

[21] . Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.