NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
SCHOOL OF MECHANICAL ENGINEERING

INTERDISCIPLINARY POSTGRADUATE PROGRAMME
“Translational Engineering in Health and Medicine”

PREDICTIVE MODELING OF MEDICATION ADHERENCE USING MACHINE LEARNING TECHNIQUES

Postgraduate Diploma Thesis

KONSTANTINA GEORGIOU

Supervisor : Dr Konstantina, S. Nikita
Professor in School of Electrical Engineer, National Technical University of Athens

Athens, October 2024

NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
SCHOOL OF MECHANICAL ENGINEERING

INTERDISCIPLINARY POSTGRADUATE PROGRAMME
"Translational Engineering in Health and Medicine"

PREDICTIVE MODELING OF MEDICATION ADHERENCE USING MACHINE LEARNING TECHNIQUES

KONSTANTINA GEORGIOU

Supervisor : Dr Konstantina, S. Nikita
Professor in School of Electrical Engineer, National Technical University of Athens

The postgraduate diploma thesis has been approved by the examination committee on *14th October 2024*

1st member	*2nd member*	*3rd member*
Dr Konstantina, S. Nikita	*Dr Georgios Stamou*	Dr Athanasios Voulodimos
Professor in School of Electrical Engineer, National Technical University of Athens	*Professor in School of Electrical Engineer, National Technical University of Athens*	*Assistant Professor in School of Electrical Engineer, National Technical University of Athens*

Athens, October 2024

.....

Konstantina Georgiou
Graduate of the Interdisciplinary Postgraduate Programme,
“Translational Engineering in Health and Medicine”,
Master of Science,
School of Electrical and Computer Engineering,
National Technical University of Athens

Abstract

This thesis aims to explore the correlation between socioeconomic factors and medication compliance in the sample of both German and Greek population based on machine learning technique. The kind of machine learning was unsupervised clustering and the algorithms utilized were K-means and hierarchical clustering.The study utilized two datasets: the German dataset, comprising 429 patients recruited from the neurology clinic at Jena University Hospital, and the Greek dataset, consisting of 81 individuals drawn from the general population. The German patients were slightly older (mean age 63.54 years) and the overall rate of adherence was higher than the Greek sample, most probably reflecting the structured clinical conditions approach of the study center. On the other hand, the Greek sample having a younger average age (average age = 54.9 years) had a broader span of adherence behaviors.

The majority of factors that were positively associated with better adherence were age, caregiver involvement in the medication preparation, unemployment and lower levels of education. In the German dataset,older patients who relied on caregivers for medication management were more likely to struggle with adherence. On the other hand, in the Greek sample, older participants who managed their own medications tended to demonstrate better adherence, emphasizing the importance of self-management in maintaining medication routines.

The K-means clustering algorithm identified distinct adherence patterns across various demographic groups, while hierarchical clustering provided a deeper understanding of the interplay between factors such as the caregiver interaction in medication preparation and socioeconomic status.

The findings underscore the critical need for designing tailored healthcare interventions that comprehensively address both demographic and socioeconomic factors. Such interventions are essential to effectively enhance medication adherence, taking into account the diverse needs and challenges present in both clinical and general population settings. By acknowledging the unique influences of age, education, employment, and caregiving roles, these strategies can foster more equitable and sustainable adherence outcomes across varied demographic groups.

Keywords:
Medication adherence, socioeconomic factors, K-means clustering, hierarchical clustering, German population, Greek population, age influence, caregiver support, self-management, employment status, education level, healthcare interventions, non-adherence, clinical population, general population, demographic factors.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά την αδερφή μου για την αδιάκοπη στήριξή της και την ενθάρρυνσή της σε κάθε βήμα αυτής της διαδικασίας. Η παρουσία της και η βοήθειά της ήταν πολύτιμη για την ολοκλήρωση αυτής της εργασίας.

Table of contents

Introduction

Medication adherence is defined as the commitment of patients to follow the prescribed regimen, encompassing adherence to timing, dosage, and frequency of the medication intake, as directed by the healthcare provider. Achieving optimal adherence is essential in pharmacotherapy, as it often directly impacts clinical outcomes. Non-adherence can lead to increased morbidity, mortality, and healthcare costs, particularly in chronic conditions where sustained medication use is critical [1]. So it is clear that through medication adherence is fundamental to improving patient health outcomes, properly managing chronic conditions, preventing complications, and enhancing the efficiency of healthcare systems by helping to lower healthcare costs[2].

The World Health Organization states that only 50% of patients with chronic diseases in developed Countries adhere to their prescribed treatment [3]. The high health and economic costs of nonadherence are a considerable challenge to society and the economy, representing a fault line in medicine. This low adherence poses a significant threat to patients, especially those with chronic diseases, as these conditions often require prolonged, even lifelong, medication to reduce morbidity or mortality. Medication is thus a pivotal component in treatment, making patient adherence to medication a crucial self-care behavior [4],[5].

Medication non-adherence is a complex issue influenced by a variety of factors, with socioeconomic barriers playing a significant role. Patients from lower-income backgrounds may face challenges such as limited access to healthcare, financial strain, and lack of health literacy, all of which contribute to lower adherence rates [6],[7]. Additionally, factors like the complexity of treatment regimens, side effects, and a lack of perceived necessity for medication also impact adherence, particularly in chronic disease management [8]. Addressing these factors through personalized care and targeted interventions is crucial to improving adherence and reducing healthcare costs.

Clustering patients based on their socioeconomic status and medication adherence can provide valuable insights for improving patient compliance. Understanding the socioeconomic factors that influence adherence, such as income, education, and healthcare access, enables healthcare providers to create more targeted interventions to address the unique challenges faced by different patient groups[9]. This approach not only allows for personalized care but also facilitates predictive analysis of adherence risks, which can lead to more efficient allocation of healthcare resources and better patient outcome. Identifying patient clusters with distinct adherence patterns helps healthcare systems implement early interventions, potentially reducing complications and preventing costly hospitalizations.

This study aims to identify distinct clusters of patients based on their socioeconomic profiles and analyze how these clusters correlate with adherence to treatment in both the German and Greek healthcare environments. By uncovering the relationship between socioeconomic status and medication adherence, the research aims to inform more effective, context-sensitive healthcare strategies to improve patient compliance and optimize resource use in these diverse populations.

Unsupervised Learning in Healthcare Data

In healthcare research, unsupervised machine learning has proven to be a promising tool for uncovering hidden connections between socioeconomic factors and medication adherence among patients. By autonomously identifying patient clusters based on shared characteristics within their socioeconomic profiles and adherence behaviors, this approach enables an exploration of how socioeconomic status influences medication adherence.[10], [11]

Several studies have utilized unsupervised clustering techniques and machine learning methods to address various challenges in healthcare. Clustering, in particular, has been employed to identify biomarkers and patterns in patient data, stratify risk groups,and uncover relationships between clinical and demographic factors, ultimately contributing to more personalized and efficient care[12],[13].

Additionally, machine learning has been increasingly applied to analyze electronic health records (EHRs) for disease prediction and diagnosis. These techniques allow for the identification of patterns within large datasets, enabling early detection of diseases and assisting healthcare providers in making more accurate diagnostic decisions. By leveraging EHR data, machine learning models can uncover subtle correlations between patient characteristics and health outcomes, significantly improving the prediction of conditions like diabetes, cardiovascular disease, and even certain cancers[14].

Unsupervised learning is a type of machine learning where models are trained using a dataset without labels, allowing them to act on data without direct supervision. This type of learning detects the underlying structure of the dataset, grouping data according to similarities and compressing information into meaningful clusters [15].

Unsupervised learning algorithms vary widely and can be selected based on the desired outcome and the dataset's characteristics. In healthcare [16], these algorithms are particularly useful for uncovering patterns and insights from complex medical data, such as socioeconomic patterns influencing adherence behavior

Some key types of unsupervised learning algorithms include:

- **K-means Clustering** is a data segmentation method that organizes information into K distinct clusters based on feature similarity. This algorithm is particularly useful for categorizing patients into various risk groups for conditions such as diabetes and heart failure, enabling more targeted healthcare interventions[17].

- **K-Nearest Neighbors (KNN):** While commonly used for classification tasks, KNN can also be applied in an unsupervised context for clustering and anomaly detection. It has proven valuable in predicting diagnostic accuracy in cancer patient data [18].

- **Hierarchical Clustering:** Hierarchical clustering builds a hierarchy of clusters using either a bottom-up (agglomerative) or top-down (divisive) approach. In the bottom-up method, each data point begins in its own cluster, with clusters progressively merged as the hierarchy grows. This technique is particularly useful for organizing complex data, such as developing taxonomies of diseases or categorizing patient profiles based on specific characteristics. Hierarchical clustering has been successfully applied in Alzheimer's disease research to classify patterns of brain atrophy[19].

- **Principal Component Analysis (PCA):** PCA is a dimensionality reduction technique that transforms complex datasets into a set of orthogonal components, capturing the maximum variance within the data. By reducing data complexity while preserving key patterns, PCA is especially valuable in fields like genomics, where high-dimensional datasets can be overwhelming. For instance, in the management of pediatric congenital adrenal hyperplasia , PCA has been used to evaluate adrenal steroid measurements and clinical markers, leading to more effective, targeted treatment strategies. PCA helps streamline large datasets into actionable insights, enhancing decision-making[20].

- **Non-negative Matrix Factorisation (NMF)**, an unsupervised machine learning technique, is applied to identify phenotype clusters in patients with Psoriatic Arthritis (PsA) based on baseline patient characteristics and clinical observations [21]. NMF is also used to characterize patterns and eventually predict the arthritis course regarding the patterns of joints involvement [22]. Unlike PCA, which allows for both positive and negative values in its components, NMF constrains the factors to be non-negative, making it particularly suited for datasets where negative values are not meaningful. Both techniques help streamline large datasets into actionable insights, enhancing decision-making in clinical and research settings.

- **Apriori Algorithm:** The Apriori algorithm is a robust tool for mining frequent itemsets and discovering association rules in large transactional datasets. In healthcare, it is used to identify co-occurring symptoms and diseases. One study applied the Apriori algorithm to investigate the network associations between diseases that frequently co-occurred in the same patient[23]. By mining these associations, both short- and long-term links between conditions were uncovered, providing critical insights into disease co-occurrence patterns and informing diagnosis and treatment strategies.

In conclusion, unsupervised learning addresses problems where data lacks labels, presenting a unique challenge as there is no definitive reference point for model evaluation. Consequently, the quality of models in unsupervised learning is typically assessed based on their outcomes, rather than through predefined human judgments.

Algorithms Used in This Study:

In this study, a range of advanced algorithms were applied to analyze both datasets.

Key methods included:

I. K-means

K-Means is particularly well-suited for the datasets in this study for the following reasons:[24]

- **Simplicity and Interpretability:** K-means is a straightforward algorithm that is relatively easy to implement and interpret, which is essential when working with healthcare data, where interpretability is a critical factor. Each cluster is defined by its centroid, representing the 'average' patient within that cluster. This allows for a clearer understanding and analysis of clusters, which can represent groups of patients with similar adherence patterns and socioeconomic backgrounds.
- **Scalability**: K-means performs efficiently on large datasets , which is important for medical and socioeconomic data, which often contain numerous variables and large numbers of patients.
- **Grouping Similar Patients:** K-means works well when the goal is to minimize variance within clusters. Since the aim is to identify groups of patients that are similar in adherence behaviors and socioeconomic characteristics, K-means is a natural choice because it groups patients in a way that minimizes the distance between patients and the center of their assigned cluster.
- **Assumption of Spherical Clusters:** K-means assumes that clusters are spherical and equally sized, which may align well with the assumption that patients with similar adherence behaviors or socioeconomic characteristics will form distinct, dense groups.

II. Hierarchical Clustering

Agglomerative clustering, a bottom-up approach within hierarchical clustering, is particularly well-suited for this dataset as it captures complex relationships between medication adherence and socioeconomic factors without assuming uniform cluster shapes or sizes. This flexibility allows it to identify naturally occurring, unevenly shaped clusters, making it ideal for healthcare data where patient profiles can vary widely [25]. Unlike other clustering methods, agglomerative clustering provides an interpretable, hierarchical structure visualized through a dendrogram, which facilitates nuanced analysis and supports decision-making in identifying distinct, layered patient groups [26]. Additionally, agglomerative clustering allows for a granular exploration of subgroups, helping researchers to detect intricate patterns within medication adherence behavior and socio-demographic factors, ultimately enabling more tailored interventions .

Theoretical Foundations of K-Means Clustering

Clustering is a core technique in exploratory data analysis, aiming to group a set of objects so that similar objects are clustered together while dissimilar objects are placed in different clusters. Unlike supervised learning, clustering is an unsupervised learning approach, which means it does not rely on predefined labels or categories. Instead, clustering algorithms seek to uncover inherent patterns and structures in the data, making it a powerful tool for revealing insights within complex datasets [27],[28].

K-means clustering is a popular partition-based clustering algorithm that aims to divide data into kkk distinct clusters. The algorithm minimizes an objective function to ensure that the variance within each cluster is reduced, typically measured by the sum of squared distances between each point and its assigned cluster centroid [29]. This process results in clusters where each data point, such as a patient's adherence behavior and socioeconomic data, is grouped with other points of similar characteristics, making it a valuable method for analyzing patient profiles.

The k-means algorithm operates in a series of iterative steps:

1. **Initialization:** The algorithm begins by randomly selecting kkk initial centroids.
2. **Assignment:** Each data point is assigned to the nearest centroid, usually measured using the Euclidean distance between the point and each centroid.
3. **Update:** The centroids are recalculated by computing the mean of all points in each cluster.
4. **Repeat:** Steps 2 and 3 are repeated until the centroids stabilize and no longer change significantly, indicating convergence.

The k-means objective function seeks to minimize the total squared distance between each data point and its assigned cluster centroid: [30]

The k-means aims to minimize the squares distance between each data point and the centroid of its assigned cluster:

$$G_{k\text{-means}}(X, C) = \Sigma_{i=1}^{k} \Sigma_{x \in C_i} ||x - \mu_i||^2$$

where μ_i is the centroid of cluster C_i

In the context of this study, each centroid represents the profile of a "typical" patient within that cluster, encapsulating both adherence patterns and socioeconomic characteristics. This interpretability makes k-means a valuable tool for identifying natural groupings in patient behaviors, which can then inform targeted healthcare strategies to enhance medication adherence and address the unique needs of each group.

Application of K-Means Clustering in Analyzing Medication Adherence and Socioeconomic Factors

The primary objective of this research is to utilize K-means clustering to analyze and categorize patients based on their medication adherence patterns and socioeconomic factors. The clustering process aims to achieve the following:

- **Prediction of Adherence Patterns:**By identifying clusters of patients with similar adherence behaviors, this research aims to pinpoint populations at higher risk of non-adherence, facilitating the development of targeted interventions and personalized strategies to improve medication adherence. Clustering patients in this way allows healthcare providers to address adherence issues with interventions tailored to specific group characteristics.
- **Analysis of Socioeconomic Impacts:**Understanding the interaction between socioeconomic status and medication adherence is essential for identifying systemic barriers and enablers of adherence. Using K-means clustering enables us to identify patterns showing how factors such as income, education level, and access to healthcare services affect adherence behaviors. This clustering method provides insights into socioeconomic impacts on health behavi**or,** highlighting areas where additional support or resources may be beneficial**.**

K-means clustering was selected for this study due to its effectiveness in partitioning datasets into distinct groups without prior labels, making it ideal for unsupervised learning tasks involving diverse patient data. This clustering approach has shown reliability in healthcare research for discovering hidden structures in data, especially in tasks focused on patient segmentation and predictive analysis of healthcare behaviors **.**

Theoretical Foundations of Agglomerative Hierarchical Clustering

Agglomerative hierarchical clustering is a bottom-up clustering technique that begins by treating each data point—here, each patient—as an individual cluster. It iteratively merges the most similar clusters based on a selected distance metric (e.g., Euclidean distance), until all points are grouped into a single overarching cluster or until a predefined number of clusters is reached. This method's ability to operate without prior knowledge of the number of clusters makes it particularly well-suited for exploratory data analysis in complex healthcare datasets [31]

Unlike partition-based algorithms like K-means, which often assume spherical and evenly sized clusters, agglomerative clustering does not require these assumptions, allowing for the creation of more flexible and complex cluster shapes. This adaptability is valuable in healthcare data, where patient adherence behaviors and socioeconomic factors often lead to non-spherical and unevenly sized clusters

The agglomerative clustering process involves several key steps:[32]

1. Initialization: Each data point begins as its own cluster.
2. Merge Clusters: In each iteration, the two clusters with the smallest distance, based on a specified distance metric, are merged.
3. Linkage Criterion: The choice of linkage criterion (e.g., Ward's method or average linkage) determines the distance calculation between clusters. Ward's method, applied in this study, aims to minimize variance between clusters, creating compact, interpretable groupings
4. Repeat: This iterative merging continues until a single cluster forms or until the desired number of clusters is reached.

The hierarchical structure of agglomerative clustering produces a dendrogram, a tree-like diagram that visually represents the clustering process. Researchers can "cut" the dendrogram at different levels, yielding varied numbers of clusters and enabling a granular exploration of the data structure

Application of Agglomerative Clustering to Analyze Medication Adherence and Socioeconomic Factors

The primary objective of this study's use of agglomerative clustering is to segment patients based on their medication adherence behaviors and socioeconomic profiles. The application of this technique is aimed at achieving two specific goals:

1. Identify Distinct Adherence Groups: By clustering patients with similar adherence patterns, this study identifies groups that may be at higher risk of non-adherence, allowing for the design of targeted interventions. These clusters provide valuable insights into behavioral trends that can inform healthcare strategies[33].
2. Explore the Role of Socioeconomic Factors: Agglomerative clustering facilitates the analysis of how various socioeconomic factors—such as age, education level, marital status, and occupation—interact with medication adherence. This method reveals natural groupings in the data, uncovering trends that may not be immediately apparent and guiding policymakers and healthcare providers in tailoring interventions to specific socioeconomic groups.

Agglomerative hierarchical clustering was chosen for its ability to uncover naturally occurring groups without requiring assumptions about the number of clusters[33]. This flexibility and interpretability make it an ideal approach for complex healthcare data, where interactions among numerous socioeconomic factors and patient behaviors create intricate patterns. By identifying these groupings, healthcare providers can develop targeted, evidence-based interventions to improve adherence, particularly among socioeconomically disadvantaged groups .

2. Methods

2.1 Description of the German dataset

Data source

The data analyzed in this study were derived from a German database that includes a wide range of demographic, socio-economic, and detailed medication adherence information [9]. The study population consisted of 429 patients diagnosed with various mental disorders. These data were collected from the Department of Neurology at Jena University Hospital in Germany.

Medication adherence was assessed using the German Stendhal Adherence to Medication Scope (SAMS) questionnaire. This questionnaire comprises 18 carefully designed questions that form a cumulative scale ranging from 0 to 72 points. A score of 0 indicates complete adherence to the prescribed medication regimen, while a score of 72 signifies total non-adherence. The SAMS questionnaire covers a broad spectrum of adherence-related behaviors, including intentional modifications to medication intake, gaps in knowledge regarding the purpose, dosage, and timing of medication, as well as instances of forgetting to take medication.

This dataset is particularly valuable as it allows for the identification of various underlying reasons and distinct clusters related to medication adherence behaviors among patients. By analyzing these data, deeper insights can be gathered into the factors influencing medication adherence and as a result targeted interventions be developed to improve adherence rates among patients.

Literature Review for the German database

The dataset used in this study has also been utilized in several other research efforts, each exploring various aspects of medication adherence in Parkinson's Disease (PD). One study [35] predicted changes in medication post-hospital discharge, highlighting the role of both intended and unintended nonadherence, with statistical analyses conducted via SPSS. Another study [36] confirmed the association between nonmotor symptoms and medication nonadherence in PD, also using SPSS for analysis. Additionally, the dataset was employed to examine the Health Locus of Control as an independent predictor of nonadherence [37], offering insights into potential intervention targets. Poor medication knowledge and its impact on health-related quality of life were investigated in another study [38], utilizing mediation analysis, while a further analysis [39] focused on identifying predictors of various types of nonadherence among elderly PD patients using descriptive statistics. These studies collectively underscore the dataset's relevance in advancing understanding of medication adherence in PD.

Statistical analysis

Data Preprocessing

Before conducting the statistical analysis, the dataset was cleaned and prepared by addressing missing values. Missing values, represented by "#" in the dataset, were replaced with NaN, and rows containing these gaps were subsequently removed to ensure the accuracy of the statistical results. As a result, 48 rows were excluded due to missing data, leaving a final dataset of 381 participants for analysis.

Statistical Analysis for age:

The age variable was analyzed, revealing a mean age of 63.54 years with a standard deviation of 15.59, indicating a diverse distribution among participants. Ages ranged from 18 to 90 years. The 25th percentile was 55 years, suggesting that a quarter of participants were 55 or younger. The median age was calculated at 68 years, meaning half of the participants were older than this value, while the 75th percentile stood at 75 years, indicating that 75% of participants were 75 or younger. This analysis underscores a predominantly older demographic, which plays a significant role in the study's outcomes.

Statistical Analysis for Adherence (SAMS score):

The adherence scores were analyzed, revealing a mean SAMS score of 6.55 and a standard deviation of 8.30, highlighting notable variability in adherence among participants. Scores ranged from 0 (fully adherent) to a maximum of 71 (severe non-adherence). The 25th percentile score was 1, suggesting that 25% of participants exhibited very low non-adherence. The median adherence score was 4, indicating that half of the participants scored below this value. The 75th percentile was recorded at 9, meaning 75% of participants had scores of 9 or lower. This distribution emphasizes a wide range of adherence behaviors, with many participants falling into the moderate adherence category, which is crucial for understanding engagement in treatment.

Investigating the Correlation Between Demographic and Socioeconomic Factors and Medication Adherence

To investigate the potential impact of demographic and socioeconomic variables on medication adherence, various statistical methods were applied. For binary data, such as gender, t-tests were conducted to assess significant differences. For categorical data with more than two groups (marital status, education, occupation), Analysis of Variance (ANOVA) was used to evaluate group variances. Regression analysis was performed for continuous numerical data to detect significant relationships.

1. T-Test for Adherence Based on Sex:
   A two-sample t-test was conducted to compare adherence scores between male and female participants. A t-statistic of 2.45 and a p-value of 0.0146 were generated, leading to the rejection of the null hypothesis. This indicated a significant difference in adherence between males and females.
2. Linear Regression for Adherence Based on Age:
   A simple linear regression analysis was conducted, with age as the independent variable and

Sum_SAMS score as the dependent variable. The R-squared value of 0.006 indicated that age did not significantly predict adherence scores, suggesting that age was not a strong predictor of adherence within this dataset.

3. ANOVA for Adherence Based on Marital Status:
   An ANOVA test was performed to compare adherence scores across marital status categories (single, married, widowed/divorced). The F-statistic was 1.27, with a p-value of 0.281, indicating no significant differences in adherence across marital status categories.
4. ANOVA for Adherence Based on Education:
   An ANOVA test compared adherence scores across different education levels, including German Realschule, German Abitur, no graduation, lower education, and university degree. The F-statistic of 0.28 and a p-value of 0.89 indicated no significant differences in adherence based on education level.
5. ANOVA for Adherence Based on Occupation:
   An ANOVA test assessed the impact of occupation (unemployed, employed, pensioned) on adherence scores. The test produced an F-statistic of 1.83 and a p-value of 0.161, indicating no significant differences in adherence based on occupation.
6. ANOVA for Adherence Based on Medication Preparation:
   An ANOVA test explored the relationship between adherence and who prepared the medication (patient, caregiver, pharmacist). The F-statistic of 6.72 and a p-value of 2.99e-05 suggested a significant impact of medication preparation on adherence, with preparation methods significantly influencing adherence scores.
7. Regression and ANOVA for Adherence Based on Medication Number:
   A regression analysis and ANOVA were conducted to examine the effect of the number of medications on adherence. The regression analysis yielded a low R-squared value of 0.027, indicating that the number of medications was not a strong predictor of adherence. The ANOVA test produced an F-statistic of 1.55 and a p-value of 0.069, suggesting a marginal effect of medication dosage on adherence but not statistically significant.

Summary of Findings:

The analysis revealed that certain factors, such as gender and the individual responsible for medication preparation, were significantly associated with adherence. However, age, education, occupation, and marital status did not show significant impacts on adherence. These insights provide valuable information on how demographic and socioeconomic factors influence adherence behavior in patients.

2.2 Description of the Greek dataset

Statistical Analysis

Data Collection and Preprocessing

The Greek dataset, developed following the structure of the German dataset, involved the administration of the translated SAMS questionnaire to the Greek population, with a total of 81 responses collected. No missing values were identified, and thus, the full dataset was retained for analysis without the need for exclusions.

Statistical Analysis for Age

The age variable was analyzed, revealing a mean age of 54.9 years with a standard deviation of 15.98, indicating a diverse range of ages among participants. Ages spanned from 23 to 85 years. The 25th percentile was 44 years, suggesting that 25% of participants were 44 or younger. The median age was calculated to be 57 years, indicating that half of the participants were older than this value. The 75th percentile stood at 66 years, showing that 75% of participants were 66 or younger. This analysis highlights a predominantly middle-aged and older demographic, which is crucial for understanding the study's adherence outcomes.

Statistical Analysis for Adherence

The adherence scores ranged from 0 to 34, with a mean of 12.54 and a standard deviation of 8.90, reflecting considerable variability among participants. The 25th percentile score was 5, indicating that 25% of participants showed low non- adherence. The median score was 12, meaning half of the participants scored below this level, while the 75th percentile stood at 18, suggesting that the majority of participants exhibited non adherence. This distribution underscores a wide range of adherence behaviors, with a substantial portion of the population falling within the moderate adherence category.

Investigating the Correlation Between Demographic and Socioeconomic Factors and Medication Adherence

To assess the influence of demographic and socioeconomic factors on medication adherence, a variety of statistical methods were applied. T-tests were used for binary variables like gender, while Analysis of Variance (ANOVA) was employed for categorical variables with more than two groups, such as marital status, education, and employment. Linear regression was used to analyze continuous variables like age.

1. T-Test for Adherence Based on Gender:
   A t-test comparing adherence scores by gender resulted in a T-statistic of -0.464 and a p-value of 0.644, indicating no significant difference in adherence between male and female participants.
2. Linear Regression for Adherence Based on Age:
   Linear regression analysis revealed a significant positive relationship between age and

adherence ($R^2 = 0.109$, $F(1, 79) = 9.668$, $p = 0.0026$), suggesting that adherence scores tended to increase with age.

3. ANOVA for Adherence Based on Marital Status:
   ANOVA analysis comparing adherence across different marital status categories (single, married, widowed/divorced) resulted in an F-statistic of 3.685 and a p-value of 0.030, showing a significant difference in adherence based on marital status.
4. ANOVA for Adherence Based on Education:
   The analysis showed significant differences in adherence scores across educational levels (F-statistic = 10.19, $p < 0.001$), indicating that education was a factor influencing adherence behavior.
5. ANOVA for Adherence Based on Employment Status:
   Significant differences in adherence based on employment status (unemployed, employed, pensioned) were found, with an F-statistic of 12.38 and a p-value of less than 0.001, suggesting employment status impacted adherence.
6. ANOVA for Adherence Based on Medication Preparation:
   ANOVA results indicated a significant difference in adherence based on who prepared the medication (F-statistic = 9.57, $p < 0.001$), showing that medication preparation responsibility influenced adherence patterns.
7. Regression and ANOVA for Adherence Based on Number of Medications:
   Regression analysis and ANOVA showed no significant relationship between the number of medications taken daily and adherence, with both analyses failing to reach statistical significance.

Summary of Findings:
The analysis identified age, marital status, education, employment, and the role in medication preparation as significant predictors of adherence. Older participants exhibited higher adherence, but the number of daily medications had no significant impact on adherence behavior. These insights highlight key factors affecting adherence in the Greek population and point to potential areas for targeted interventions.

2.3 Adherence Categorization

To facilitate the interpretation of adherence behavior within each cluster, participants were categorized into three adherence groups based on their SAMS score:

- **Fully adherent** (SAMS score = 0)
- **Moderate non-adherent** (SAMS score between 1 and 10)
- **Non-adherent** (SAMS score > 10)

In the current study, the Stendal Adherence to Medication Score (SAMS) was utilized to categorize patient adherence into three distinct groups. The cut-off point of 10 for clinically significant non-adherence was supported by previous research [34], which identified a SAMS score of 0.80 (equivalent to 80% adherence) as a valid threshold for predicting adverse outcomes such as

hospitalization across chronic diseases. This cut-off was adapted to align with the specifics of the dataset, ensuring a clinically meaningful categorization of adherence. This categorization was subsequently applied to analyze the distribution of adherence behavior within the identified clusters, providing a clearer understanding of how adherence varies across different socio economic groupings

3. Machine Learning

3.1 K-means Clustering analysis for the German dataset

To explore the relationship between socioeconomic factors and medication adherence, K-Means Clustering was applied. This method allowed for the segmentation of participants into distinct clusters, providing insights into patterns of adherence behavior across different socioeconomic groups.

Data Preparation
Before performing the clustering analysis, the dataset was prepared by selecting relevant socioeconomic variables, including age, sex, marital status, education, occupation, medication preparation method, and medication dosage. The total adherence score, derived from the sum of SAMS items 1 to 18, was incorporated as a key feature to ensure that patients were grouped based on their overall adherence behavior. The data preparation process involved the following steps:

- Handling Categorical Variables: Categorical features such as sex, marital status, education, occupation, and medication preparation method were transformed into numerical format using one-hot encoding. This conversion allowed categorical data to be represented in a binary format, making it suitable for the clustering algorithm.
- Standardization: To ensure equal contribution of all features in the clustering process, the data was standardized using StandardScaler. This step normalized the data by transforming it to have a mean of 0 and a standard deviation of 1, a necessary procedure for machine learning models like K-Means that are sensitive to varying data scales.

Determining the Optimal Number of Clusters
The optimal number of clusters was identified using the elbow method, which evaluates the within-cluster sum of squares (WCSS) across various numbers of clusters, ranging from 1 to 10. The "elbow point" in the plot of WCSS values against the number of clusters was determined, indicating the most suitable number of clusters. Based on this analysis, three clusters were selected as the optimal solution.

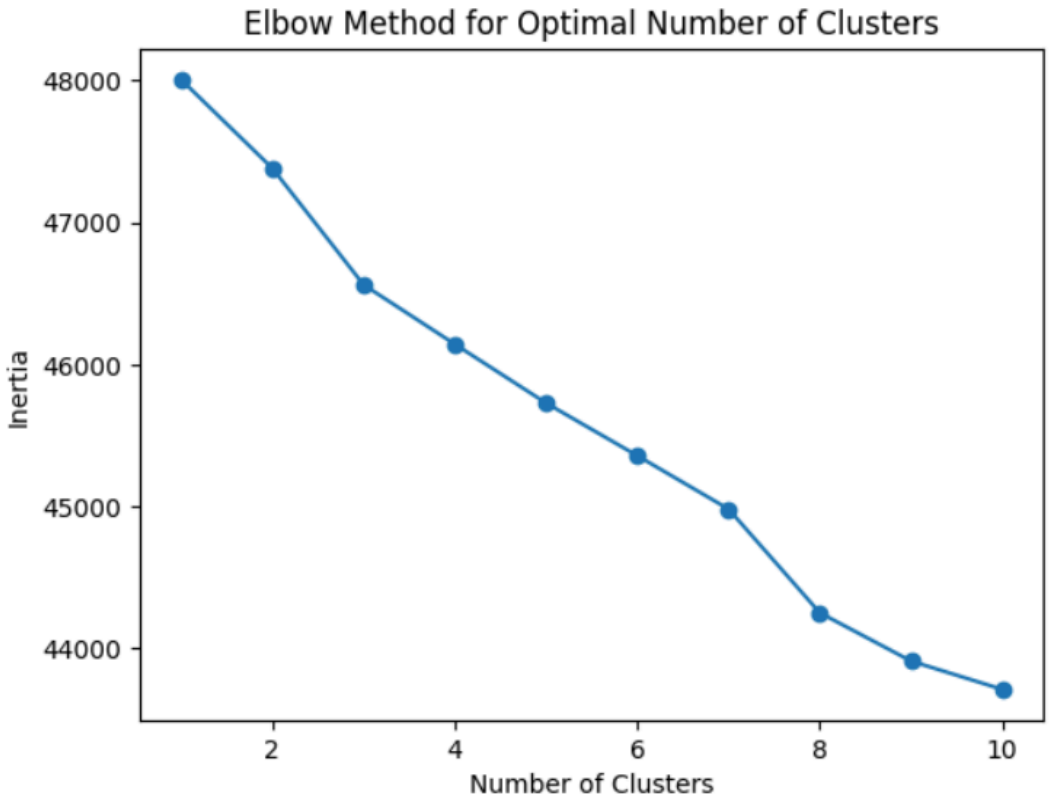

Figure 1. Elbow method to determine the optimal number of clusters

Grid Search for Fine-Tuning:
A Grid Search was conducted to fine-tune the parameters of the K-Means algorithm. This process involved testing a range of hyperparameters, including different values for the number of clusters, initialization attempts, and random seed values. The GridSearchCV function was used to exhaustively search through the parameter grid, evaluating model performance for each combination. The best-performing parameters were identified as follows:

- Number of clusters: 5
- Number of initializations: 20
- Random state: 10

This optimization ensured a more precise K-Means clustering model by improving clustering performance and minimizing inertia while exploring various configurations.

Additionally, silhouette analysis was performed to further validate the selection of the optimal number of clusters. Silhouette scores, which assess how well an object fits within its own cluster compared to others, were calculated for different cluster configurations. The configuration with the highest average silhouette score supported the selection of two clusters, indicating better-defined cluster separation.

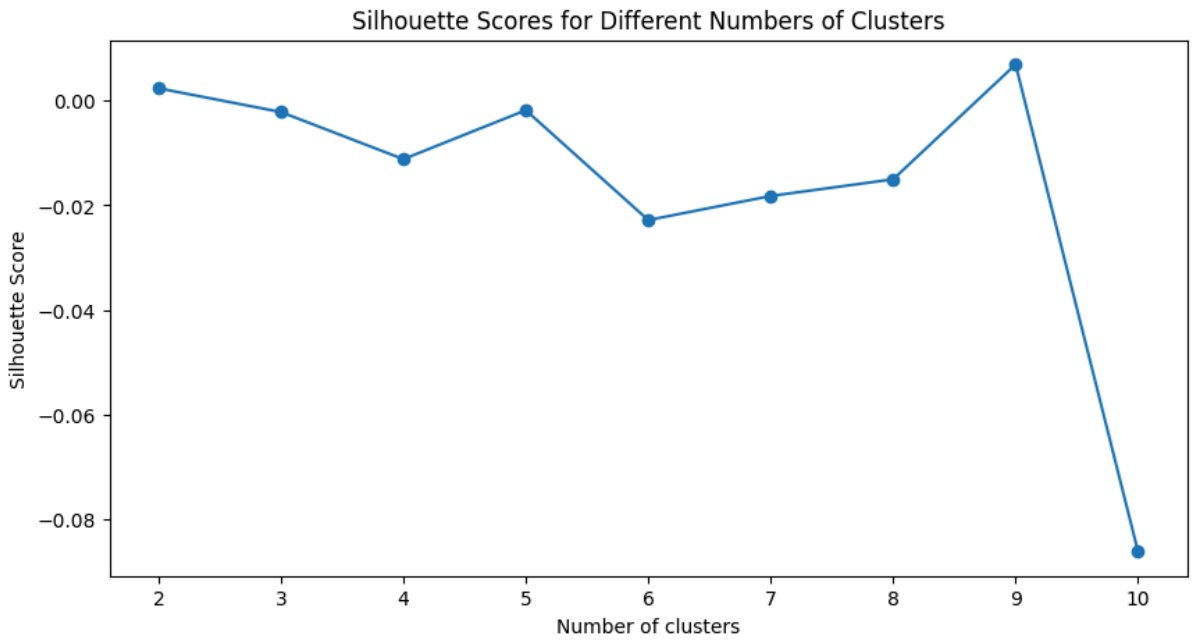

Figure 2. Silhouettes scores to determine the optimal number of clusters

K-Means Clustering Implementation

After the optimal cluster count was determined, the K-Means algorithm was applied with five clusters based on the fine-tuning process and silhouette score analysis. Each participant in the dataset was assigned to one of these clusters according to their socioeconomic characteristics and adherence score. This allowed for the identification of distinct groups, each representing different patterns of adherence behavior within a socioeconomic context.

Principal Component Analysis (PCA) was also utilized for dimensionality reduction, facilitating the visualization of the clusters in a two-dimensional space. By plotting the first two principal components, the spatial distribution of the clusters was effectively illustrated, highlighting the separation of clusters based on the combination of socioeconomic factors and adherence behaviors.

Cluster Analysis Results

Through K-means clustering, the dataset was segmented into five clusters, derived from key socioeconomic variables and medication adherence. The elbow method confirmed that five clusters provided the best balance between data compactness and separation. These clusters were then analyzed in relation to the SAMS adherence categories, which were grouped into three classifications: fully adherent, moderately non-adherent, and non-adherent.

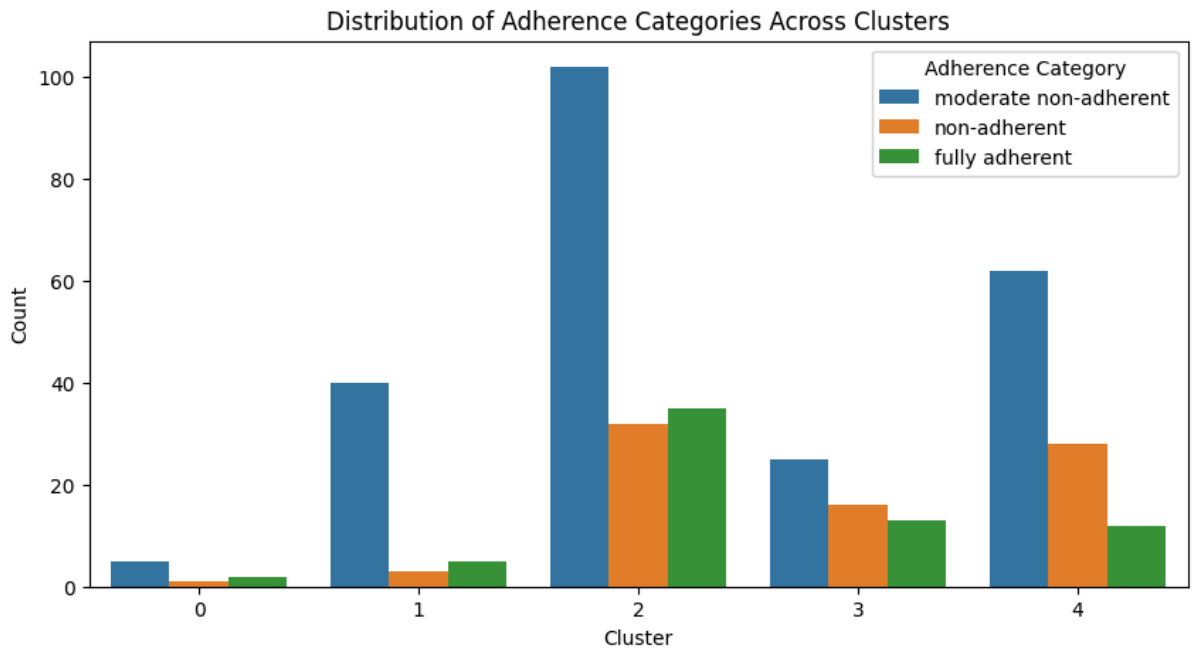

Figure 3. Distribution of adherence categories across clusters in german dataset

As illustrated above, Cluster 2 was found to contain the largest proportion of participants, with a notable number classified as having moderate non-adherence (102) and a significant number categorized as fully adherent (35). In contrast, Cluster 0 comprised the fewest individuals, with only a small representation of fully adherent (2) and non-adherent participants (1). Clusters 3 and 4 displayed a more balanced distribution across all three adherence categories, highlighting variability in medication-taking behaviors within these groups.

Cluster Visualization with PCA:

Principal Component Analysis (PCA) was utilized to reduce the dimensionality of the data, enabling the creation of a 2D scatter plot (Figure No#) for visualizing the clusters. The PCA plot revealed that the clusters were generally well-separated, although some overlap was observed, particularly between Clusters 1, 2, and 4, suggesting similarities in socioeconomic characteristics among these clusters.

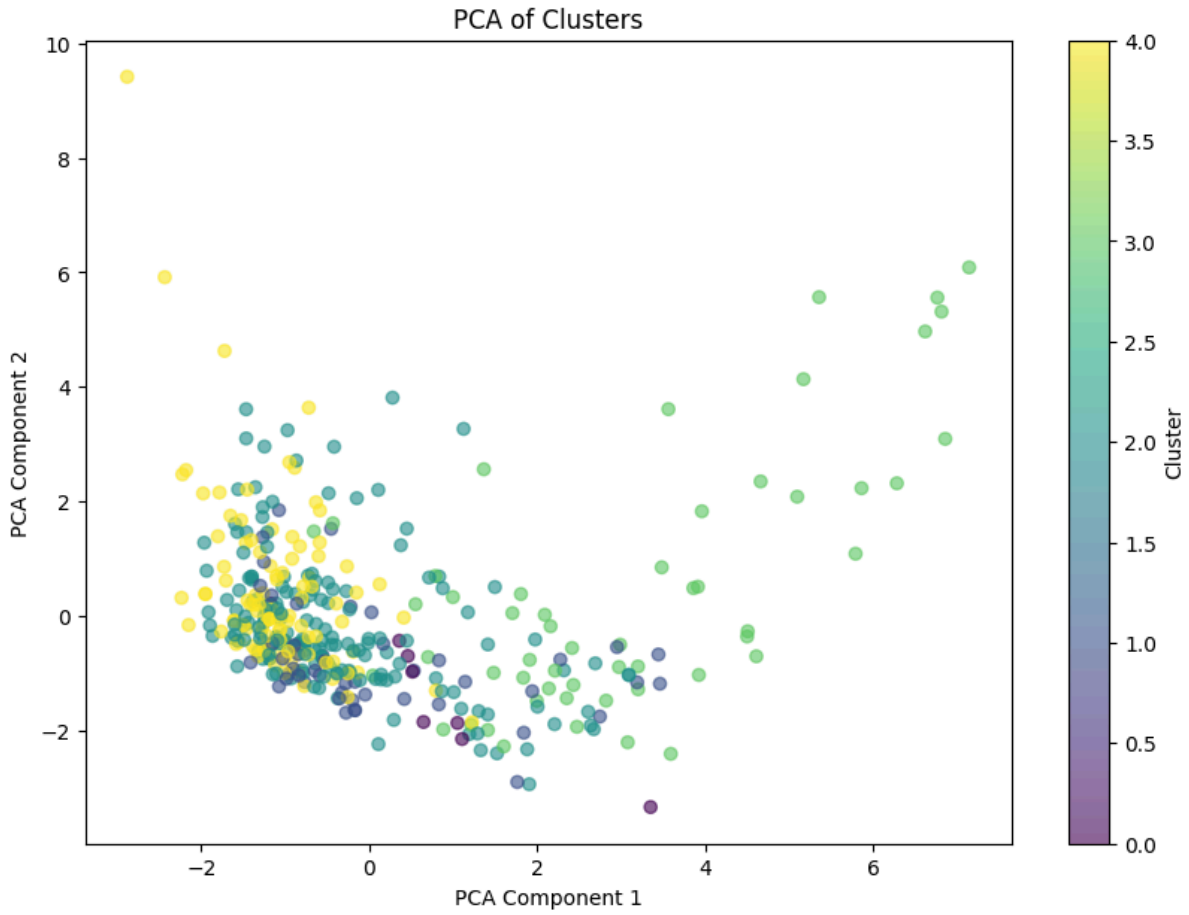

Figure 4. Principal Component Analysis of clusters in german dataset

The clustering analysis revealed distinct patterns in medication adherence, with certain clusters demonstrating a higher tendency toward non-adherence. These findings suggest that underlying socioeconomic factors play a role in influencing adherence behaviors, and the identified clusters provide valuable insights for developing targeted interventions tailored to different patient groups.

A comprehensive analysis of the identified clusters was conducted. Summary statistics for each cluster were calculated, focusing on the numeric variables. Both the mean and standard deviation were computed to compare the average values and variability across the clusters, offering a clearer understanding of the socioeconomic characteristics and adherence behaviors within each group.

Below there is a table with the statistics finding of every cluster.

Cluster Summaries

Cluster 0:

- Age: Participants in this cluster had an average age of 55.5 years, representing a middle-aged demographic.
- SAMS Score: This group exhibited medium SAMS score (Mean SAMS score = 4.13), indicating moderate adherence to medication.
- Sex: The distribution was evenly split between males and females, with each gender representing 50% of the cluster.
- Marital Status: Participants were balanced between those who were married and those who were single.
- Education: The educational level was more evenly distributed, with 25% of the participants holding a university degree.
- Occupation: A relatively high employment rate was observed, with 62.5% of participants currently employed.
- Medication Management: Most participants in this cluster managed their own medications independently, without the need for caregiver assistance.

Cluster 1:

- Age: This cluster had an older average age of 62.9 years.
- SAMS Score: Participants demonstrated strong adherence, with a low mean SAMS score of 2.2, suggesting generally high adherence.
- Sex: The majority of participants were female (66.7%).
- Marital Status: A high proportion of participants were married, with 75% reporting being in a marital relationship.
- Education: There was a relatively even distribution of education levels, though 37.5% of participants had lower education levels.
- Occupation: The majority of participants were pensioners (77%), reflecting the cluster's older demographic.
- Medication Management: Most participants (72.9%) self-managed their medications, while a smaller portion (14.6%) relied on caregiver assistance.

Cluster 2:

- Age: The average age in this cluster was 65.3 years, representing an older cohort.
- SAMS Score: This cluster had a higher mean SAMS score of 7.1, indicating issues with medication adherence and a tendency toward non-adherence.
- Sex: The gender distribution was relatively balanced, though slightly more females (52.7%) were represented in this group.

- Marital Status: Most participants were married (69.8%).
- Education: Nearly half of the participants (45.6%) had completed secondary education.
- Occupation: The majority of participants were retired, with 81% reporting themselves as pensioned.
- Medication Management: A significant proportion of participants (79.3%) managed their medications independently, with some assistance from caregivers.

Cluster 3:

- Age: This was the youngest cluster, with an average age of 40.7 years.
- SAMS Score: The average SAMS score was 6.85, reflecting moderate non-adherence to medication.
- Sex: A large majority of participants in this cluster were male (72.2%).
- Marital Status: The majority of participants were single, with 81.5% reporting this status.
- Education: Participants were diverse in terms of educational background, with 31.5% having completed secondary education and 20% not having completed any formal education.
- Occupation: Over half of the participants were employed (53.7%), though a significant proportion (27.8%) were unemployed.
- Medication Management: The rate of self-management was lower in this group (66.6%), with a higher reliance on caregiver assistance for medication adherence.

Cluster 4:

- Age: This was the oldest group, with an average age of 73.6 years.
- SAMS Score: This cluster exhibited the highest non-adherence, with a mean SAMS score of 7.7.
- Sex: The majority of participants were male (76.5%).
- Marital Status: The majority of participants were married, with 85.3% reporting this status.
- Education: This cluster had the highest proportion of participants with a university degree (50.9%).
- Occupation: Almost all participants in this cluster were pensioned (95.1%).
- Medication Management: Most participants managed their own medications, with some receiving assistance from caregivers or pharmacists.

Key Observations Across Clusters:

Cluster 0 (Balanced Demographics, Moderate Adherence):
This cluster, characterized by middle-aged participants and a balanced sex ratio, exhibited moderate adherence challenges. A significant portion of the group was employed, and the majority managed their medications independently.

Cluster 1 (Older, Good Adherence, High Marriage Rates):
Participants in this older cluster, predominantly female and mostly married, showed few adherence issues. The majority were pensioned, with high levels of medication adherence reported.

Cluster 2 (Older, Moderate Adherence Issues):
This group, with a slightly older demographic and balanced gender representation, displayed moderate adherence issues. The majority were pensioned, and challenges in medication adherence were noted.

Cluster 3 (Younger, Male-Dominated, Moderate Non-Adherence):
Comprising the youngest participants, this male-dominated cluster was primarily single and employed. Despite these characteristics, moderate non-adherence was observed, potentially linked to life-stage factors.

Cluster 4 (Oldest, High Non-Adherence, High Education):
This cluster, the oldest and predominantly male, included a high proportion of participants with advanced education. However, it also exhibited the highest non-adherence scores, with nearly all participants being pensioned.

Key Summary:
Adherence trends across clusters highlight that older populations (Clusters 1, 2, and 4) show greater variability in medication adherence, with a notable increase in non-adherence among the oldest group. Younger participants, particularly in Cluster 3, also demonstrated moderate non-adherence, potentially influenced by challenges such as unemployment or other life-stage factors. These findings underscore the need for targeted adherence interventions tailored to both age-related and socioeconomic factors.

3.2 K-means Clustering analysis for the Greek dataset

Handling Categorical Variables Using One-Hot Encoding

Several categorical variables were present in the dataset, including gender, marital status, education level, employment status, medication preparation responsibility, and the total number of daily medications. These categorical variables required transformation before applying clustering algorithms, such as K-Means, which operate exclusively on numerical data. To address this, One-Hot Encoding was utilized to convert each categorical variable into a binary representation (0 or 1). This encoding process generated new binary variables for each category within the features, ensuring compatibility with the K-Means algorithm.

Elbow Method for Determining the Optimal Number of Clusters

The K-Means algorithm was implemented to identify distinct groupings within the dataset. However, one critical aspect in applying K-Means is determining the optimal number of clusters (k), which significantly impacts clustering outcomes. To ascertain the appropriate number of clusters, the Elbow Method was employed. This method evaluates the within-cluster sum of squares (WCSS) for different values of k, and the "elbow point" on the plot helps to identify the most appropriate number of clusters. The elbow point suggested that four clusters would offer an optimal balance between compactness and separation.

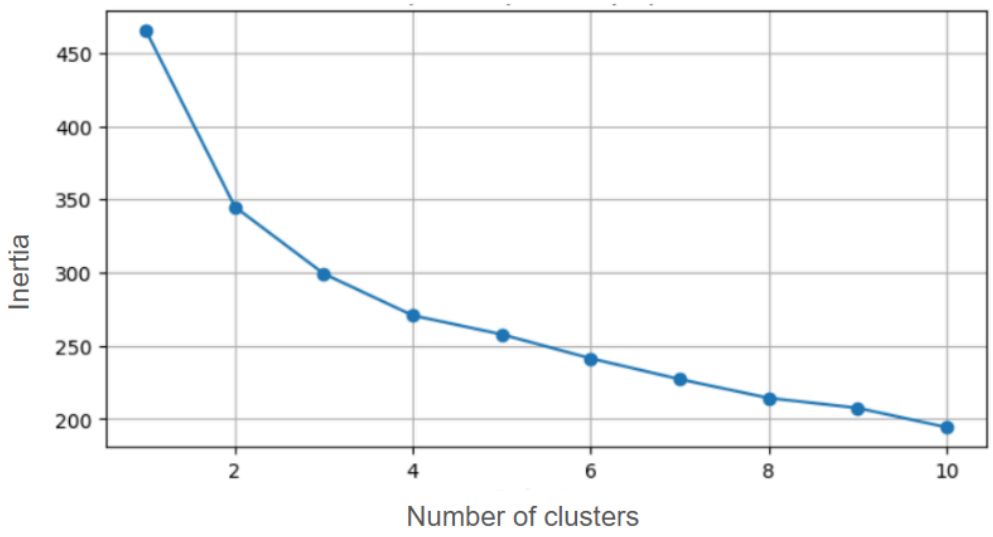

Figure 5. Elbow method in greek dataset

Grid Search for Fine-Tuning

Subsequent to determining the number of clusters, a Grid Search was conducted to fine-tune the hyperparameters of the K-Means algorithm. This process involved testing a variety of hyperparameter combinations, including different values for the number of clusters, initialization attempts, and random seed values. GridSearchCV exhaustively evaluated these combinations to identify the best-performing parameters. The optimal configuration identified:

- Number of clusters: 5
- Number of initializations: 50
- Random state: 42

Clustering Methodology

The K-Means algorithm was employed with four clusters (k=4), as identified through the Elbow Method. The algorithm iteratively assigned data points to the nearest cluster centroid, updating the centroid positions based on the mean of the points within each cluster. This iterative process continued until convergence was achieved. To visualize the resulting clusters, Principal Component Analysis (PCA) was applied to reduce the dimensionality of the data, transforming it into two principal components. The PCA scatter plot illustrated a clear distinction between the clusters, with each cluster represented by a different color, demonstrating separation based on socioeconomic factors and medication adherence.

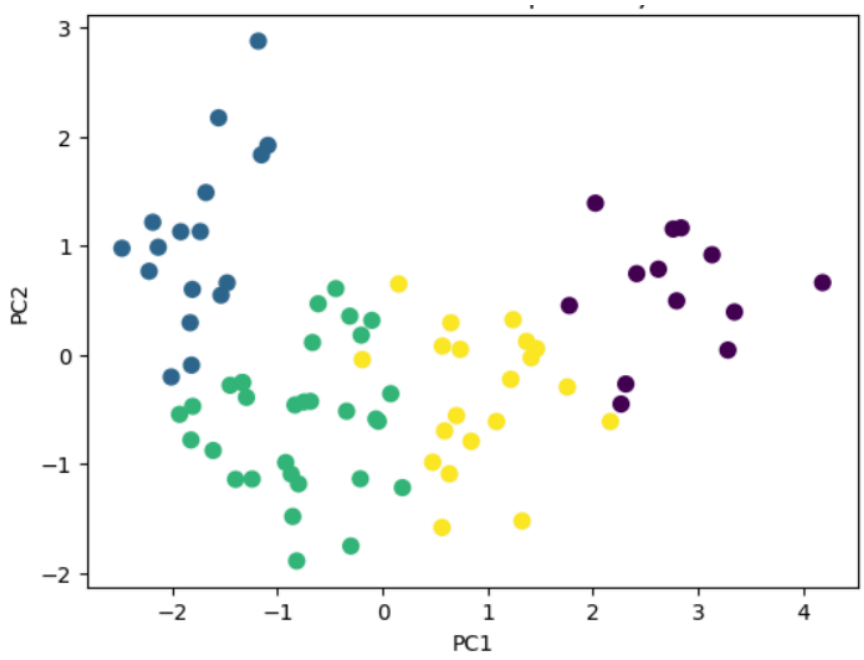

Figure 6. PCA clusters in greek dataset

Cluster Evaluation Using Silhouette Score

The effectiveness of the clustering solution was evaluated using the Silhouette Score, a metric that measures how well each point matches its own cluster compared to neighboring clusters. The score ranges from -1 to 1, with values closer to 1 indicating better-defined clusters. In this analysis, the average Silhouette Score for four clusters was 0.18, suggesting moderate separation between clusters. Additionally, a silhouette plot was generated, providing further insight into the clustering structure by illustrating the silhouette coefficients for each point in every cluster.

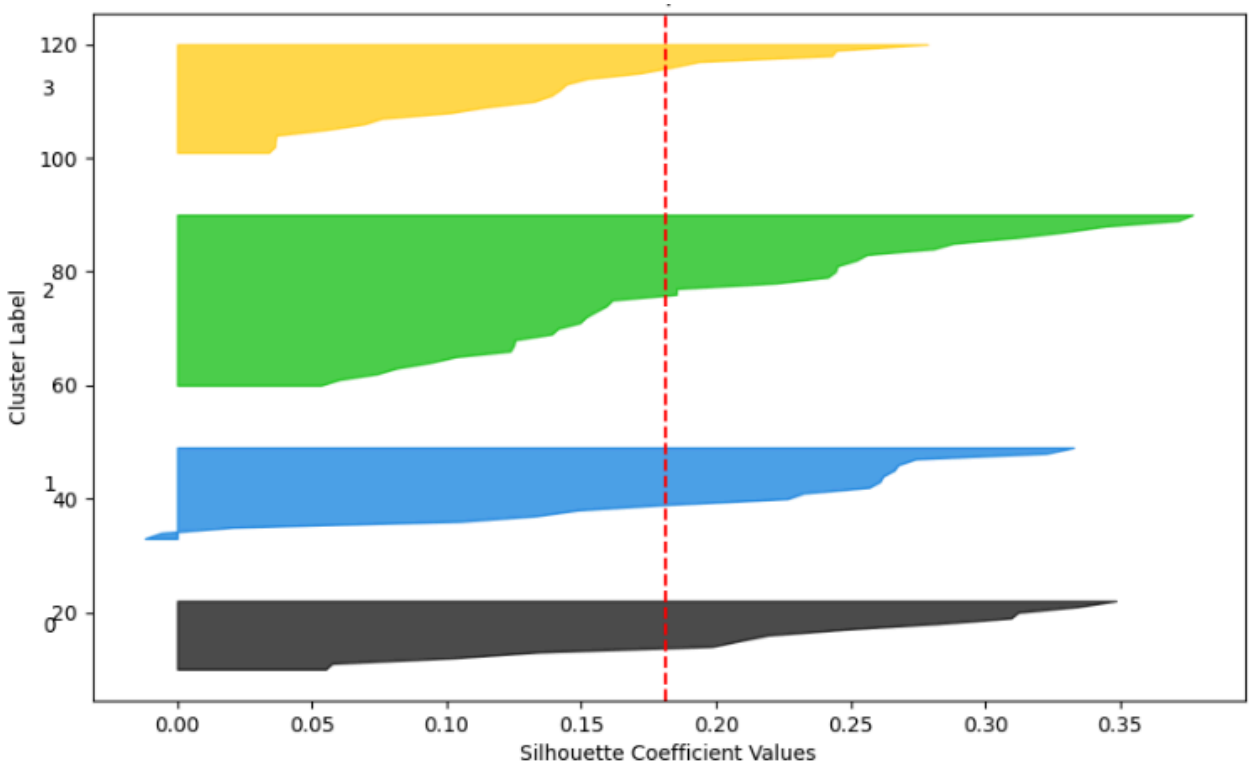

Figure 7. Silhouette scores of clusters in greek dataset

Cluster Summaries:

Cluster 0 (Elderly, High Medication Usage, Non-Adherent)
This cluster comprises an elderly population with an average age of 72.15 years. Participants exhibited the highest medication usage, averaging 5.85 medications daily. However, the SAMS score (25.46) indicated significant non-adherence. Demographically, most participants were female (61.5%) and widowed or divorced (76.9%). Nearly half (46.2%) had completed high school, and all were retired. Notably, 53.8% relied on family for medication management, a factor likely contributing to the observed non-adherence.

Cluster 1 (Young Adults, Low Medication Usage, Moderate Non-Adherence)
The youngest group, with an average age of 30.88 years, reported low medication usage, with participants taking an average of 1.06 medications daily. The SAMS score (10.53) reflected moderate non-adherence. This cluster was predominantly female (76.5%) and single (94.1%). A high percentage (70.6%) had completed their education, and most were employed (70.6%). Participants were generally independent in medication management, with 70.6% managing their medications themselves.

Cluster 2 (Middle-Aged, Low Medication Usage, Adherent)
Cluster 2 represented middle-aged participants, with an average age of 54.13 years. Medication usage was low, averaging 1.94 medications daily. The SAMS score (6.87) indicated full to moderate adherence, highlighting better adherence compared to Clusters 0 and 1. This cluster had more males (61.3%), and 64.5% were married. Only a small percentage (19.4%) had completed secondary education. The cluster had a high employment rate (87.0%), with most individuals self-managing their medications (90.3%), contributing to better adherence.

Cluster 3 (Older Adults, Moderate Medication Usage, Non-Adherent)
This older cluster had an average age of 65.30 years and reported moderate medication usage, with participants taking 3.45 medications daily. The SAMS score (14.65) indicated non-adherence, although less severe than in Cluster 0. The cluster exhibited equal male and female representation, and 80.0% were married. Educationally, 40.0% had completed high school, with some participants pursuing further education. The majority (75.0%) were retired, and most (85.0%) self-managed their medications, although adherence remained a challenge.

Summary of Findings
The clustering analysis revealed significant differences in medication adherence and management across the four identified clusters. Cluster 0, characterized by an elderly population dependent on family for medication management, demonstrated the highest non-adherence levels. In contrast, Cluster 2, comprising middle-aged, employed individuals who largely managed their medications independently, exhibited the best adherence. Younger adults in Cluster 1 faced moderate non-adherence but managed their medications independently, while older adults in Cluster 3 experienced moderate non-adherence despite self-management. These findings emphasize the importance of addressing adherence challenges in elderly populations and promoting self-management strategies across all age group.

Hierarchical clustering

In this study, hierarchical clustering using the agglomerative approach was applied to analyze the dataset, which includes factors related to medication adherence and socioeconomic characteristics. The methodology followed several key steps, detailed below, to ensure the effective clustering of the data and the extraction of meaningful patterns.

3.3 Hierarchical clustering for the german dataset

Data Preprocessing

The dataset consisted of both categorical and numerical variables. As a first step, categorical variables were converted into numerical form through One-Hot Encoding. This process was essential for transforming categorical variables such as sex, marital status, education, occupation, and medication preparation responsibility into a binary matrix suitable for clustering. The target variable for this analysis, medication adherence, was represented by the Sum_SAMS items 1 till 18 column, which was included alongside socioeconomic features.Next, the data was standardized using StandardScaler, ensuring that numerical variables like age and the adherence score were on the same scale.

Linkage Calculation and Dendrogram Construction

The Ward linkage method was chosen to compute the linkage matrix for hierarchical clustering. This method minimizes the variance within clusters, ensuring that the clusters remain as homogeneous as possible, which is important when analyzing complex datasets like healthcare data. The linkage matrix was used to create a dendrogram, which visualized how the clusters were formed and provided insight into where an appropriate cut-off point for clustering should be drawn.

In this case, a cut-off line was drawn at a dendrogram height of 7 to generate an initial grouping of the data. This threshold was selected after visually inspecting the dendrogram to capture a balance between forming distinct groups and avoiding over-segmentation of the data.

Agglomerative Clustering Implementation

After determining the appropriate number of clusters from the dendrogram, Agglomerative Clustering was applied to the dataset. Various cluster numbers were tested, ranging from 2 to 5 clusters, to find the most appropriate solution. The final solution of 4 clusters was chosen for the German dataset. This decision was based on the optimal combination of interpretability and clustering performance metrics.

The 4-cluster solution allowed for clear separation between different groups, offering a nuanced understanding of patient profiles in terms of medication adherence and socioeconomic factors. Each cluster represents a distinct group of patients, enabling more tailored analyses of patient characteristics and behaviors.

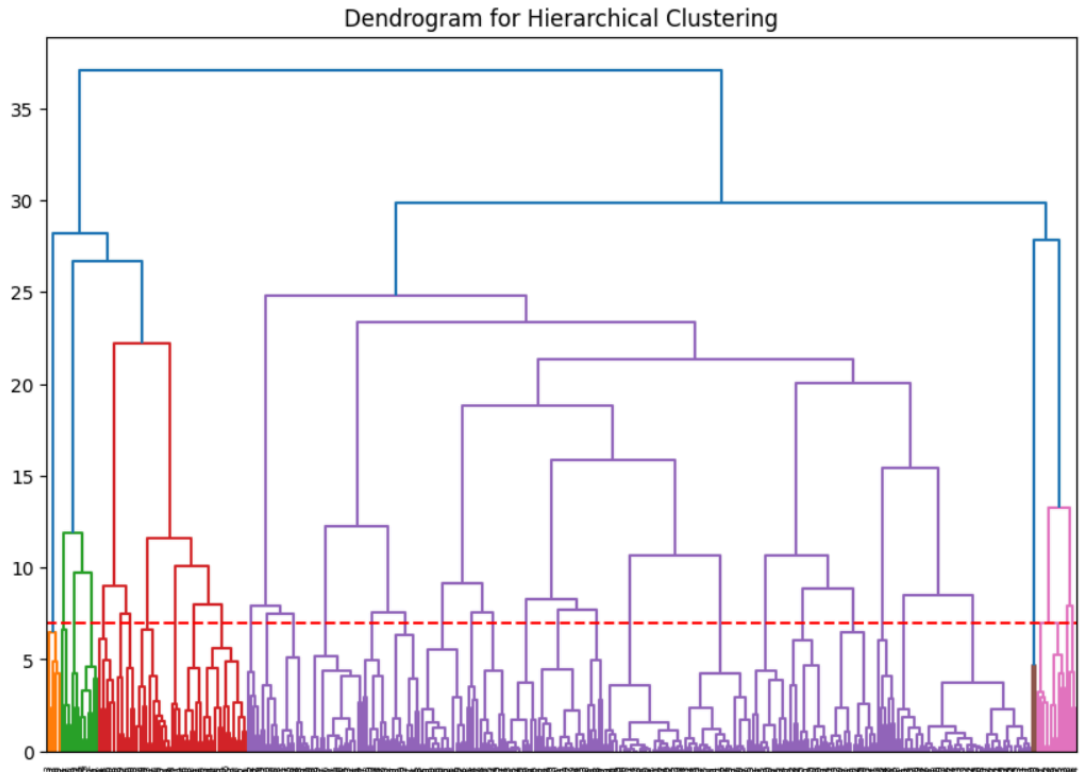

Figure 8. Dendrogram in agglomerative clustering in german dataset

Dimensionality Reduction and Visualization

To visualize the clusters in two dimensions, Principal Component Analysis (PCA) was performed, reducing the dataset to two principal components. This facilitated clearer visualization of the clusters, making it easier to interpret the relationships between groups.

A scatter plot was generated from the PCA-transformed data, where the clusters showed adequate separation. This confirmed that the hierarchical clustering method had successfully segmented the dataset into meaningful groups, despite the high dimensionality of the original data.

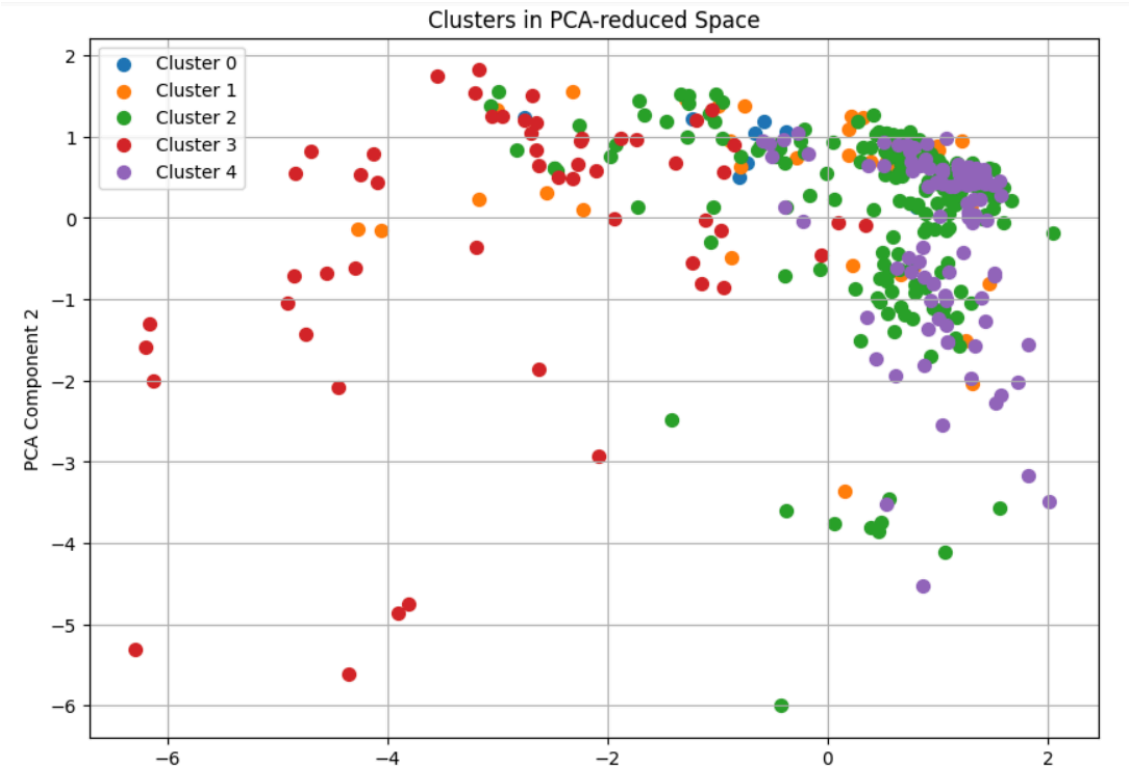

Figure 9. Principal Component Analysis of clusters in german dataset

Evaluation of Clustering Performance

To assess the quality of the clustering, two performance metrics were computed: the Silhouette Score and the Calinski-Harabasz Index. These metrics were calculated for cluster solutions ranging from 2 to 5 clusters to determine the most appropriate number of clusters for the dataset.

Number of Clusters	Silhouette Score	Calinski-Harabasz Index
2	0.2739	51.71
3	0.2955	46.63
4	0.2891	45.88
5	0.2998	47.43

- Silhouette Score favors 5 clusters (0.2998), suggesting that the clusters are more clearly defined with this number of clusters.
- Calinski-Harabasz Index strongly supports 2 clusters (51.71), indicating that the data separation is stronger with fewer clusters.

Although 5 clusters provide a better Silhouette Score, 2 clusters offer the strongest overall separation. If you prefer well-defined clusters in terms of internal cohesion, 5 clusters might be the best choice. However, if you're aiming for more distinct clusters based on group separation, 2 clusters might be more suitable.Given the moderate increase in the Silhouette Score for 5 clusters, it could be worth choosing 5 clusters for a more granular analysis, depending on your goal.

Clusters Analysis

Cluster 0 (Younger, Single, Moderate Adherence):

- Age: The average age is 44.3 years, making this group younger than the others.
- Gender: 55.1% of the participants are male.
- Marital Status: 84.1% are single, while only 5.8% are widowed or divorced.
- Education: 27.5% have a lower education level (e.g., German Hauptschule), and 20.3% did not complete formal education. Just 13.0% hold a university degree.
- Occupation: 29.0% are pensioned, and 26.1% are unemployed.
- Medication Management: 79.7% of the participants manage their medication independently, with only 1.4% relying on others for medication preparation.
- Adherence: The average adherence score is 5.2 on the Sum_SAMS scale, reflecting moderate adherence.

This cluster is characterized by a relatively younger demographic with a slight majority of males. The majority of participants are single, and a significant portion is either unemployed or pensioned. Despite some holding university degrees, a large percentage have lower educational qualifications. Most individuals manage their medication independently, and their adherence levels are moderate, likely influenced by their employment and social status.

Cluster 1 (Older, Low Adherence, Dependent on Caregivers):

- Age: The participants in this group are older, with an average age of 63.1 years.
- Gender: 53.3% are male.
- Marital Status: 26.7% are single, and 33.3% are widowed or divorced.
- Education: None of the participants in this cluster have a university degree, and 33.3% have a lower education level.
- Occupation: A high 73.3% of the participants are pensioned, while 20.0% are unemployed.
- Medication Management: All participants (100.0%) rely on caregivers for medication management, with none managing their medication themselves.
- Adherence: The average adherence score is 14.9, indicating non adherence.

Comprising older participants, this cluster shows a low level of medication adherence, with all participants receiving support from caregivers for medication management. A large proportion of this group is pensioned, and a significant number are widowed or divorced. While educational attainment is generally low, the cluster's dependency on external medication support highlights its vulnerability and the need for consistent healthcare intervention.

Cluster 2 (Older, Independent, Moderate Adherence):

- Age: With an average age of 68.5 years, this cluster represents an older demographic.
- Gender: 57.9% male.
- Marital Status: No participants are single, but 17.9% are widowed or divorced.
- Education: 23.8% have a lower education level, while 30.7% hold a university degree, representing one of the more highly educated clusters.
- Occupation: 86.2% are pensioned, with none being unemployed.
- Medication Management: A large proportion (78.9%) manage their medication independently.
- Adherence: The average adherence score is 6.5, suggesting moderate adherence.

This group, with an older demographic and a balanced gender distribution, displayed moderate adherence issues. The majority of individuals are pensioned and manage their medication independently, despite complex health needs. A relatively high proportion holds university degrees, suggesting that education contributes to their ability to handle medication on their own, although adherence challenges still persist.

Cluster 3 (Younger, Adherence, No Systematic Medication Use):

- Age: The average age is 38.8 years, placing this cluster in the younger demographic.
- Gender: 40% of participants are male.
- Marital Status**:** 40% are single, and 20% are widowed or divorced.
- Education: Educational levels are varied, with 20% holding a university degree and 20% having a lower education level.
- Occupation: 20% are pensioned, and none are unemployed.
- Medication Use: Participants in this cluster do not report systematic medication use, indicating an absence of consistent reliance on medical treatments.
- Adherence: The average adherence score is 1.0, reflecting high adherence to health-related behaviors despite the lack of regular medication use.

Cluster 3 is composed of younger individuals who exhibit no systematic medication use. This group displays diverse educational attainment and a relatively balanced gender distribution. While the absence of consistent medication use may suggest a generally healthy population, it also highlights potential vulnerabilities regarding future health risks or preventive care.

Cluster 4 (Oldest, High Education, Pharmacist-Dependent):

- Age: This cluster has the oldest participants, with an average age of 76.0 years.
- Gender: 54% are male.
- Marital Status: None are single, but 50.0% are widowed or divorced.
- Education: 50.0% have a lower education level, and another 50.0% hold a university degree, indicating high educational attainment in this cluster.
- Occupation: 100.0% are pensioned, with no unemployed participants.
- Medication Management: Almost all participants rely on pharmacists to manage their medication.
- Adherence: The average adherence score is 10.5, reflecting moderate adherence.

This is the oldest cluster, with an equal gender split and the highest levels of educational attainment. Despite their advanced age, participants in this cluster show moderate adherence scores, and 100% rely on pharmacists for medication management. Most individuals are pensioned and widowed or divorced, indicating that this group may face challenges related to aging and social isolation, though their reliance on pharmacists suggests strong healthcare support systems are in place.

3.4 HierArchical Clustering for the Greek dataset

Data Preprocessing

The dataset included both categorical and numerical variables. To handle the categorical features, **One-Hot Encoding** was used, transforming each categorical variable into a binary matrix. This encoding method was chosen to convert categorical responses such as gender, family status, education, and employment into a format suitable for clustering algorithms, which require numerical input. Subsequently, **standardization** was applied to the numerical variables (age, GR_Sum_SAMS, and total daily medication intake).

Linkage Calculation and Dendrogram Construction

To visualize the clustering structure, the **Ward linkage method** was employed to compute the linkage matrix, which was then used to generate a **dendrogram**. The Ward method was selected for its ability to minimize the variance within clusters at each step of the clustering process. This method ensures that the resulting clusters are as compact as possible, which is particularly important when analyzing healthcare-related data where interpretability and precision are key. The dendrogram provided a visual summary of the hierarchical clustering process, illustrating the distances at which clusters merged. A cut-off line was drawn at a height of 15 to yield a suitable number of clusters for further analysis.

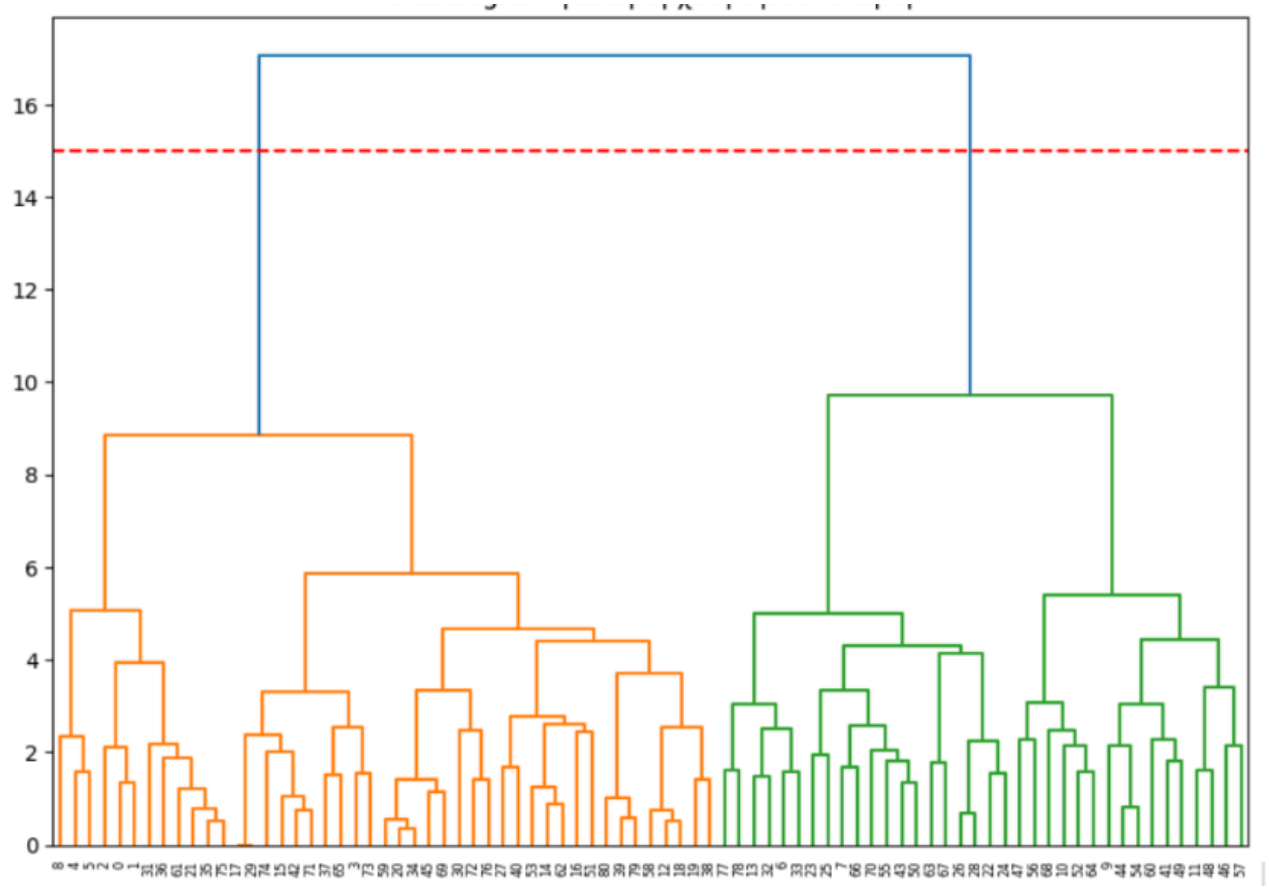

Figure 10. Dendrogram in agglomerative clustering in greek dataset

Agglomerative Clustering Implementation

Following the dendrogram analysis, **Agglomerative Clustering** was applied to the dataset, with the number of clusters set to three. The decision to use **two clusters** was based on a balance between interpretability and performance. While the dendrogram indicated potential cluster separations, the three-cluster solution was specifically chosen as it aligns with practical considerations related to healthcare interventions. The division into three distinct groups allows for more targeted analysis of patient behaviors and socioeconomic factors while maintaining a level of simplicity that facilitates actionable insights. Each cluster represents a distinct patient profile, reflecting variations in medication adherence and related socioeconomic characteristics.

Dimensionality Reduction and Visualization

To facilitate visualization, **Principal Component Analysis (PCA)** was performed, reducing the high-dimensional dataset into two principal components. This enabled a clear two-dimensional representation of the clusters. The scatter plot revealed that the clusters were well-separated, indicating that the agglomerative clustering had successfully segmented the data into meaningful groups.

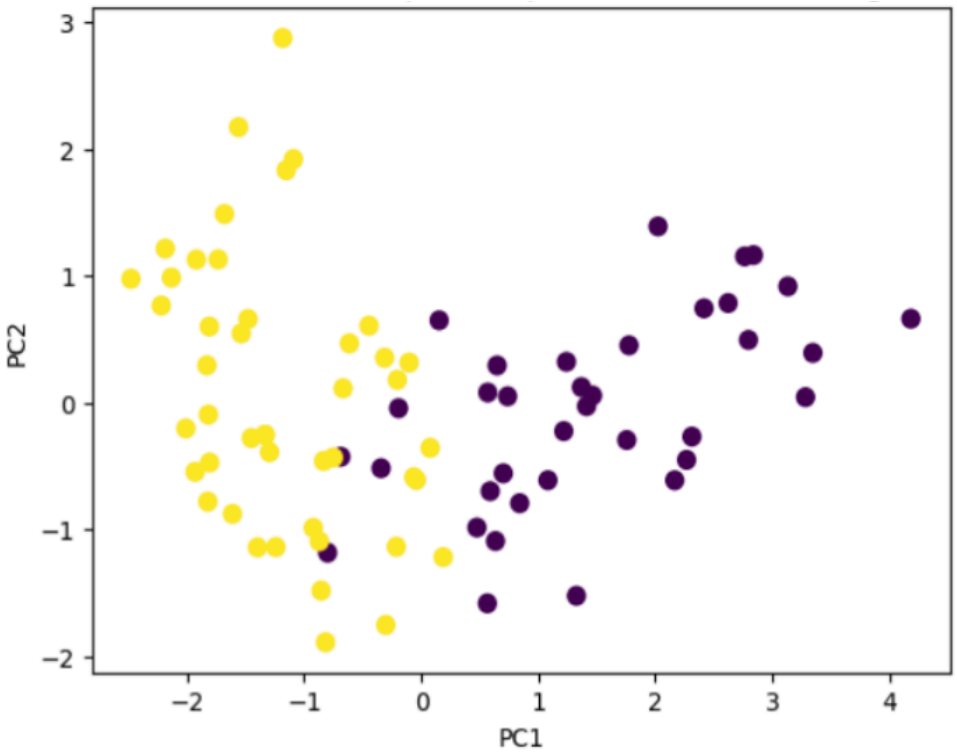

Figure 11. Principal Component Analysis of clusters in greek dataset

Evaluation of Clustering Performance

To assess the quality of the clusters, both the **Silhouette Score** and the **Calinski-Harabasz Index** were computed for solutions ranging from 2 to 5 clusters.

Number of Clusters	Silhouette Score	Calinski-Harabasz Index
2	0.2550	34.17
3	0.1839	26.17
4	0.1721	24.03
5	0.1436	20.54

Based on the Table above ,the **2-cluster solution** was selected because it provides the highest **Silhouette Score (0.255)**, indicating better-defined and more distinct clusters compared to other solutions. Moreover, the **Calinski-Harabasz Index (34.40)** is also the highest for 2 clusters, suggesting stronger separation between clusters and greater overall cohesion within each cluster. The metrics for 3, 4, and 5 clusters show diminishing returns in both the Silhouette Score and Calinski-Harabasz Index, further supporting the decision to opt for a simpler, more robust 2-cluster solution.

Cluster Analysis

Cluster 0 (Older, Retired, Moderate Medication Use):

- Age: The average age is approximately 66.2 years, making this group older than many other clusters.
- Gender: Approximately 61.1% of the participants are female, while 38.9% are male.
- Marital Status:About 50% of participants are married. A significant proportion (around 38.9%) are widowed or divorced, and the remaining 11.1% are single.
- Education:Approximately 22.2% have only completed primary education, while 38.9% are high school graduates (either technical or general).Around 30.6% hold higher education degrees (university or tertiary education).
- Occupation: A vast majority (around 83.3%) are retired, indicating that they are no longer in the workforce, while a small percentage (approximately 8.3%) are unemployed.
- Medication Management:About 63.9% of participants manage their medication independently, with 19.4% relying on nursing staff and 16.7% having a family member or caregiver prepare their medications.
- Medication Use:Participants report taking an average of approximately 5.5 medications daily, reflecting a typical scenario for older adults managing multiple health conditions.

- Adherence: The adherence score on the Sum_SAMS scale varies, with average 7.8.

This cluster is characterized by an older demographic with a slight majority of females. Most participants are retired and manage multiple medications, indicative of their age and potential health conditions. The educational background is diverse, with a significant portion having lower educational qualifications. While many individuals manage their medications independently, there are notable numbers relying on family or nursing staff. Overall, medication adherence is moderate, influenced by the complexity of their regimens and health literacy levels.

Cluster 1 (Younger, Employed, High Education, Moderate Adherence)

- Age: The average age is approximately 40.4 years, indicating that this group is younger compared to other clusters.
- Gender: Approximately 48.9% of the participants are female, showing a balanced distribution.
- Marital Status: About 57.8% of participants are single, with 26.7% being married, and a notable portion (15.6%) identified as widowed or divorced.
- Education: A significant majority (around 68.9%) have completed higher education (university or tertiary education), while 11.1% are high school graduates and another 11.1% have vocational training.
- Occupation: An overwhelming majority (approximately 84.4%) are employed, reflecting active participation in the workforce, while a small percentage (8.9%) are unemployed.
- Medication Management: About 66.7% of participants manage their medication independently, with a minority (4.4%) relying on family members for assistance and 2.2% depending on pharmacists.
- Medication Use: Participants report taking an average of approximately 2.0 medications daily, which is indicative of a generally lower need for medication among younger adults.
- Adherence: The adherence score on SAMS around 9.0, showing moderate adherence.

This cluster is characterized by a younger demographic, with a nearly equal gender distribution. Most participants are actively employed and possess high educational qualifications, reflecting their health literacy. They manage their medications independently, with a relatively low average medication intake, indicating fewer health conditions compared to older groups. Overall, medication adherence is moderate to high, influenced by their education and proactive health management approaches.

4. Conclusion

The comprehensive analysis of clustering results across the German, Greek, and SAMS datasets using both K-Means and Hierarchical clustering techniques highlights significant variations in demographic characteristics, medication adherence behaviors, and healthcare engagement across different clusters. Each dataset and clustering method has provided unique insights into participants' medication adherence and lifestyle factors, which reflect the diverse challenges and opportunities in managing health outcomes across different age groups, marital status, educational background, and social support networks.

The Greek dataset encompasses a broader population, showcasing various ages and medication management practices. In contrast, the German dataset focuses on older, hospitalized individuals, emphasizing the importance of caregiver support and higher adherence levels among those reliant on external assistance. These insights underline the need for tailored healthcare interventions that consider demographic factors to enhance medication adherence across diverse populations. Healthcare strategies should be adaptable, ensuring that interventions are not only disease-specific but also aligned with individual demographics, social circumstances, and support systems.

Statistical Analysis of the Greek and German Datasets.

The statistical analysis of the Greek and German datasets revealed key differences and similarities in the demographic characteristics and adherence behaviors of the participants in each study. These insights provide a foundation for understanding the factors that influence medication adherence in two distinct populations: one comprised of ordinary people in Greece and the other consisting of neurology patients in German hospitals.

Age Distribution and Demographic Characteristics.

The age distribution between the Greek and German datasets demonstrated distinct differences. In the Greek dataset, the mean age was 54.9 years (SD = 15.98), with ages ranging from 23 to 85 years. The median age was 57 years, with 25% of participants aged 44 or younger, and 75% aged 66 or younger. This distribution points to a predominantly middle-aged and older population, although the presence of younger participants was also noted.
In contrast, the German dataset had a higher mean age of 63.54 years (SD = 15.59), with ages ranging from 18 to 90 years. The median age was 68 years, with the 25th and 75th percentiles at 55 and 75 years, respectively. This reflects a predominantly older population in the German dataset, which is expected given that the data were collected from neurology patients in hospitals, a group that generally tends to skew older. The older demographic in the German dataset likely plays a significant role in adherence behaviors, as older patients are often more likely to encounter health challenges requiring consistent medication adherence [21].

Medication Adherence

Adherence behaviors, as measured by the SAMS score, varied considerably between the two datasets. In the Greek dataset, the mean adherence score was 12.54 (SD = 8.90), with scores ranging from 0 to 34. The median score was 12, and the 75th percentile was 18, indicating that a majority of participants experienced some level of non-adherence, with a significant portion of the population falling into the moderate adherence category. The Greek population's adherence scores displayed substantial variability, reflecting a wide range of medication-taking behaviors that may be influenced by a combination of socioeconomic factors and personal responsibility in managing medication.

In the German dataset, the mean adherence score was 6.55 (SD = 8.30), with scores ranging from 0 to 71. The median adherence score was 4, and the 75th percentile stood at 9, indicating that 75% of the participants had low to moderate levels of non-adherence. This lower mean score compared to the Greek dataset suggests that the German participants, despite being a clinical population, exhibited relatively better adherence overall. This could be attributed to the fact that the German participants were hospital-based neurology patients, who may have received more structured support or monitoring for their medication adherence. Additionally, these participants may have had more direct interaction with healthcare providers, which could have positively influenced their adherence behaviors.

German K-Means Clustering Analysis

The German K-Means clustering results yielded five distinct clusters with varying demographic and behavioral patterns. Cluster 0 represents a middle-aged demographic with an average age of 55.5 years and a moderate adherence to medication, with a mean SAMS score of 4.13. This cluster has a balanced gender distribution and an even split in marital status between single and married individuals. Most participants manage their medications independently, reflecting a strong self-management orientation.

Cluster 1, on the other hand, consists of an older group with an average age of 62.9 years. This cluster showed high adherence to medication with a low mean SAMS score of 2.2. A majority of participants in this cluster were female, and a substantial proportion were married. While most individuals self-managed their medications, a notable segment relied on caregiver support. Cluster 2, consisting of older participants with an average age of 65.3 years, displayed higher non-adherence issues, as indicated by a SAMS score of 7.1. Despite balanced gender distribution, the majority were married, and a significant number managed their medications independently but required some caregiver support. Cluster 3 represents the youngest group, with average age 40.7 years. IT showed moderate medication adherence with a SAMS score 6.85 and a dominant male demographic. Most participants were single, with varied educational backgrounds, which influenced their reliance on caregivers and self-management practices. Cluster 4 comprised the oldest group with an average age of 73.6 years. This cluster displayed the highest non-adherence, reflected by a SAMS score of 7.7. A large proportion of participants were male and married, with most relying on pharmacists or caregivers for medication support.

German Hierarchical Clustering Analysis

The hierarchical clustering approach in the German dataset provided further insights into participants' medication adherence. Cluster 0 featured younger, single individuals with an average age of 44.3 years. Despite a large percentage being single, most managed their medications independently with only 1.4% relying on external assistance. The average adherence score for this cluster was 5.2 on the Sum_SAMS scale, indicating moderate adherence. Cluster 1, composed of older individuals with an average age of 63.1 years, displayed low adherence to medication. All participants in this cluster depended entirely on caregivers, highlighting their vulnerability and reliance on external support for medication management. The adherence score was particularly low at 14.9, reflecting a critical need for healthcare interventions and systemic support. Cluster 2 included older individuals with an average age of 68.5 years, showing moderate adherence at an average score of 6.5. Most participants in this group managed their medications independently, and a significant number held university degrees. Cluster 3 consisted of younger individuals with an average age of 38.8 years and no systematic medication use, reflecting a cluster where adherence was less about pharmacological intervention and more about proactive health behaviors. Lastly, Cluster 4 highlighted older participants with an average age of 76.0 years, showcasing high educational attainment and dependence on pharmacists for medication management, with an adherence score of 10.5 on the Sum_SAMS scale.

Greek K-Means Clustering Analysis

The Greek dataset analysis using K-Means clustering further illustrated the relationship between age, medication usage, and adherence. Cluster 0, the elderly group, had an average age of 72.15 years and exhibited high medication usage, taking an average of 5.85 medications daily. Despite this high medication load, the SAMS score of 25.46 indicated significant non-adherence. This cluster, predominantly female, relied heavily on family members for medication management, which could contribute to the observed non-adherence rates. Cluster 1 consisted of young adults with an average age of 30.88 years, taking an average of 1.06 medications daily. The SAMS score of 10.53 indicated moderate non-adherence, with most participants being single and female. Most individuals managed their medications independently, showcasing a relatively high adherence despite lower medication demands. Cluster 2, composed of middle-aged individuals with an average age of 54.13 years, showed good adherence levels with a SAMS score of 6.87, supported by high self-management rates and low medication needs. This cluster highlighted the benefits of employment and education in fostering independent medication management. Cluster 3 in the Greek K-Means dataset featured older adults averaging 65.3 years, with moderate medication use and a SAMS score of 14.65. Despite a balanced gender distribution, most participants relied on self-management, even though adherence challenges persisted.

Greek Hierarchical Clustering Analysis

The hierarchical clustering analysis of the Greek dataset also revealed significant insights into medication adherence. Cluster 0 consisted of older, retired individuals with an average age of 66.2 years. This group managed an average of 5.5 medications daily and displayed a moderate adherence score of 7.8. Educational backgrounds varied, with 38.9% having high school education and about 30.6% holding higher education degrees. Most participants managed their medications independently, but a significant portion still required support from family members or nursing staff. Cluster 1, composed of younger employed individuals with an average age of 40.4 years, demonstrated a high educational attainment (around 68.9% with higher education) and a robust workforce participation rate (approximately 84.4%). These factors contributed to a moderate adherence score of 9.0 on the Sum_SAMS scale, with 66.7% managing their medications independently. The employment status and educational background appeared to play a crucial role in enhancing self-management practices and adherence to medication regimens.

Patterns of Medication Adherence

Medication adherence varies significantly across different clusters in the German and Greek datasets, shaped by factors such as age, social support, education, healthcare interactions, and family involvement. By examining the patterns of adherence, we can identify groups with good, moderate, and low adherence, each with distinct characteristics and influencing factors.

Good Adherence

In both the German and Greek datasets, good medication adherence is primarily seen in groups that benefit from strong support systems, education, and healthcare interactions.

In the German dataset, Cluster 1, with an average age of 62.9 years, demonstrates high adherence due to substantial reliance on caregivers and healthcare professionals. Participants in this cluster receive consistent support that ensures proper medication intake, highlighting the crucial role of external assistance.

In the Greek dataset, Cluster 2, which includes middle-aged individuals averaging around 54.13 years, exhibits good adherence due to independent medication management and employment stability. This group self-manages their medications effectively and benefits from a balance of education and job engagement, which fosters discipline and health responsibility.

Participants with good adherence often have higher levels of education, stable social relationships, and structured interactions with healthcare providers. They are typically able to manage their medications independently or rely on professionals and family members who prioritize consistent healthcare routines.

Moderate Adherence

Groups with moderate adherence usually show a balance between self-management and external support, where the presence of family, healthcare interactions, and social networks can vary in their influence.

In the German dataset, Cluster 2 (average age 65.3 years) demonstrates moderate adherence, where most participants manage their medication independently, though some still require occasional support from caregivers.

In the Greek dataset, Cluster 1, which includes younger adults around 30.88 years of age, reports a moderate adherence score of 10.53. Despite low medication usage, this group maintains adherence through self-management and active engagement with healthcare routines, driven by higher education levels and professional stability.

Individuals in these groups exhibit adherence influenced by employment status, family engagement, and social interactions, balancing self-reliance with occasional external assistance. These factors foster a discipline of self-management while ensuring medication adherence through support networks and healthcare interactions.

Low Adherence

Low adherence is typically observed in groups where social support is limited, education levels are lower, or healthcare interactions are less structured. These groups experience challenges due to a lack of consistent healthcare interactions and limited support systems.

In the German dataset, Cluster 4, consisting of older participants averaging 73.6 years, shows high non-adherence with a mean SAMS score of 7.7. These individuals primarily rely on pharmacists and external caregivers but still face challenges in maintaining consistent adherence due to age-related cognitive and physical limitations.

In the Greek dataset, Cluster 0 (elderly with an average age of 72.15 years) experiences significant non-adherence. Participants in this group take an average of 5.85 medications daily, but the SAMS score of 25.46 highlights non-adherence. Factors contributing to this low adherence include heavy reliance on family members for medication management, cognitive impairments, and logistical challenges in healthcare delivery.

Low adherence groups are characterized by minimal healthcare interactions, reliance on familial support rather than healthcare professionals, and challenges in cognitive and social aspects of medication management. Often, these individuals face difficulties due to family dynamics, lack of healthcare infrastructure support, and limited health literacy.

5. Discussion

This study employed both K-means and hierarchical clustering methods to analyze medication adherence behaviors in Greek and German populations. The insights gained from the analysis underscore the importance of tailoring healthcare interventions to the specific demographic, socioeconomic, and healthcare contexts of different populations. Both methods offered unique strengths in understanding the factors influencing adherence, but each also revealed certain limitations that warrant consideration.

Insights from Clustering Analysis

The clustering analysis highlighted key distinctions in medication adherence behaviors between the Greek and German datasets, driven by age, socioeconomic factors, healthcare settings, and medication management practices. These findings align with prior research emphasizing the influence of demographic and contextual factors on adherence, but they also reveal nuances specific to these two populations.

Age-Related Trends:
Age emerged as a significant determinant of adherence in both datasets, with older participants generally demonstrating higher non-adherence, particularly in the German hospital-based dataset. This trend may reflect the greater complexity of medication regimens among older populations, compounded by reliance on caregivers for medication management. In the Greek dataset, however, older adults who managed their own medications tended to exhibit better adherence, suggesting that independence and self-efficacy play a critical role in adherence outcomes.

Role of Socioeconomic Factors:
Socioeconomic variables, including employment, education, and marital status, had differing impacts on adherence across the two datasets. In the Greek dataset, these factors were more pronounced, with employed, middle-aged participants showing moderate non-adherence likely due to the demands of balancing work and healthcare responsibilities. In contrast, the German dataset, collected from a more controlled hospital environment, showed less influence of socioeconomic factors, as the standardized care provided in hospitals may mitigate these disparities.

Medication Management Practices:
The role of self-management versus caregiver involvement was a pivotal factor in adherence behaviors. In the Greek dataset, clusters with higher rates of self-management demonstrated better adherence outcomes. Conversely, in the German dataset, clusters with higher reliance on caregivers, particularly among older adults, exhibited lower adherence. This finding underscores the importance of equipping caregivers with the knowledge and resources needed to support effective medication adherence, particularly in clinical settings where patients often depend on external assistance.

Comparison of Clustering Methods: K-Means vs. Hierarchical Clustering

When comparing K-Means and Hierarchical clustering methods, it becomes evident that hierarchical clustering offers deeper insights into medication adherence patterns across both the German and Greek datasets.

In the German dataset, hierarchical clustering provides more granular details of caregiver interactions, social support networks, and healthcare dependencies, which are crucial for understanding adherence among older adults. It effectively identifies the dependencies that shape adherence behaviors, such as the reliance on caregivers and pharmacist interactions.

In the Greek dataset, hierarchical clustering again proves superior, capturing complex social dynamics within family interactions and support networks. It reveals the interactions that drive adherence patterns among elderly individuals and younger adults, highlighting the influence of family support and healthcare engagement on medication management.

While K-Means is computationally efficient and scalable, it often oversimplifies social interactions and contextual factors, which are better captured by hierarchical clustering methods. In both datasets, hierarchical clustering excels at revealing social and familial interactions that directly impact adherence, making it the preferred method for deeper qualitative and quantitative healthcare research.

In conclusion, the comparative analysis of medication adherence across the German and Greek datasets, using both K-Means and hierarchical clustering methods, highlights the critical role of demographic, social, and contextual factors in medication adherence. Good adherence is observed in groups with strong healthcare support, stable social interactions, and higher educational attainment, where medication self-management and professional interactions play significant roles. Moderate adherence is characterized by a balance of self-management and occasional external assistance, where social networks and employment stability contribute significantly to adherence discipline. Low adherence is seen in groups with insufficient healthcare interactions, limited support systems, and lower cognitive or educational engagement, which pose challenges for consistent medication intake.Hierarchical clustering emerges as a superior method for capturing adherence patterns across both datasets, thanks to its ability to reveal intricate social support networks, caregiver interactions, and healthcare dependencies.

Implications for Healthcare Interventions

The findings of this study have several implications for improving medication adherence across diverse populations:

Tailored Interventions for Older Adults:
Older adults face adherence challenges due to age-related cognitive, physical, and social constraints. These challenges were evident in the German dataset (Cluster 4) and the Greek dataset (Cluster 0), where elderly participants experienced significant non-adherence. The personas representing older adults often highlight a reliance on family members for medication management, cognitive impairments, and logistical challenges in healthcare delivery.Healthcare interventions should focus on enhancing caregiver support and collaboration with healthcare professionals. Caregiver Training Programs would empower family members and healthcare aides with the skills to support medication routines, ensuring proper intake, managing side effects, and maintaining consistent healthcare interactions.Simplifying Medication Regimens through tools like pill dispensers, pre-packaged medications, and automated reminders can address confusion and reduce errors. Additionally, telehealth services provide accessible healthcare consultations, ensuring that patients receive continuous support without frequent hospital visits. Community healthcare initiatives should also involve health navigators who can assist older adults in managing their medications and connecting with healthcare professionals.

Promoting Self-Management:
Self-management initiatives, as seen in the Greek dataset (Cluster 2, middle-aged individuals averaging 54.13 years) and Cluster 1 in the German dataset, show the effectiveness of independent medication management and professional stability. These personas demonstrate that higher adherence is often associated with education, stable employment, and proactive healthcare engagement.Healthcare interventions should focus on educational programs that promote self-management skills, teaching patients about medication schedules, side effects, and adherence strategies. Workshops and online resources can provide accessible knowledge about health responsibility and medication discipline.Support Groups and Peer Networks can also play a significant role in reinforcing self-management habits. Encouraging patients to share experiences and strategies can foster discipline and accountability. For example, peer groups for individuals with chronic illnesses could include activities focused on sharing medication management strategies, discussing challenges, and reinforcing healthy routines.

Addressing Socioeconomic Barriers:
Socioeconomic factors were identified as significant contributors to medication adherence challenges in the Greek dataset, where groups such as Cluster 0 (elderly with an average age of 72.15 years) often experience difficulties due to financial constraints and employment instability. The personas representing individuals with lower income and unstable employment show that financial pressures, job insecurity, and lack of healthcare infrastructure support contribute to non-adherence.Healthcare interventions should include Financial Assistance Programs, subsidizing medication costs for lower-income individuals and ensuring that

medications remain accessible. Governments and healthcare organizations should implement employment flexibility policies that allow workers to attend medical appointments and maintain consistent healthcare interactions.Community-Based Support Groups offer emotional and practical support by providing medication reminders, healthcare consultations, and community health initiatives. These initiatives can include health fairs, free screenings, and local workshops that provide essential health education and resources.

Leveraging Healthcare Systems:
The German dataset highlights the importance of **structured healthcare interactions**, as seen in Cluster 1, where high adherence was facilitated by consistent support from healthcare providers and caregivers. This suggests that **expanding hospital-based benefits to community healthcare settings** can improve adherence across larger populations.Community Health Programs should offer regular health check-ups, consultations, and workshops at local healthcare centers, ensuring continuous interactions between patients and healthcare providers. These community-based interactions should be accessible and consistent, allowing healthcare professionals to monitor adherence, adjust treatment plans, and address potential issues before they escalate. Telehealth Services also offer scalable solutions for maintaining engagement with healthcare professionals. Regular video consultations and virtual check-ups can bridge geographical gaps, ensuring patients have access to professional advice, medication reviews, and adherence strategies without requiring long trips to healthcare facilities.

Gender-Sensitive Approaches:
Gender differences observed in adherence behaviors suggest the need for tailored approaches that address the unique challenges faced by men and women. For example, younger men in the Greek dataset exhibited lower adherence, which may indicate the need for more engaging and accessible healthcare outreach targeted at this demographic.

Limitations and Future Directions

While this study provided valuable insights, certain limitations should be acknowledged. The Greek dataset represented a general population, while the German dataset was hospital-based, potentially limiting the generalizability of the findings to other settings. Future research should aim to include more diverse populations, including rural and underserved communities, to provide a more comprehensive understanding of adherence behaviors. Additionally, while clustering methods revealed important patterns, combining these approaches with advanced machine learning techniques could yield even more precise insights into the predictors of adherence. Incorporating longitudinal data could also enhance our understanding of how adherence behaviors evolve over time and in response to interventions.

In conclusion, the comparative analysis of Greek and German datasets revealed that adherence behaviors are shaped by a complex interplay of demographic, socioeconomic, and healthcare-related factors. While self-management generally improves adherence, older populations with complex medication needs require additional support. Socioeconomic stability and caregiver involvement also play critical roles, emphasizing the need for tailored interventions that address these contextual factors. By leveraging the strengths of both K-means and hierarchical clustering, this study provides a foundation for designing targeted healthcare programs aimed at improving medication adherence across diverse populations.

References

[1] M.T. Brown and J.K. Bussell, "Medication adherence: WHO cares?" *Mayo Clin. Proc.*, vol. 86, no. 4, pp. 304-314, Apr. 2011, doi: 10.4065/mcp.2010.0575. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/21389250.

[2] M.C. Roebuck, J.N. Liberman, M. Gemmill-Toyama, and T.A. Brennan, "Medication adherence leads to lower health care use and costs despite increased drug spending," *Health Aff. (Millwood)*, vol. 30, no. 1, pp. 91-99, Jan. 2011, doi: 10.1377/hlthaff.2009.1087. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/21209444.

[3] M.T. Brown and J.K. Bussell, "Medication adherence: WHO cares?" *Mayo Clin. Proc.*, vol. 86, pp. 304-314, 2011, doi: 10.4065/mcp.2010.0575.

[4] Z.U. Rehman, A.K. Siddiqui, M. Karim, H. Majeed, and M. Hashim, "Medication non-adherence among patients with heart failure," *Cureus*, vol. 11, e5346, 2019.

[5] T.M. Ruppar, P.S. Cooper, D.R. Mehr, J.M. Delgado, and J.M. Dunbar-Jacob, "Medication adherence interventions improve heart failure mortality and readmission rates: Systematic review and meta-analysis of controlled trials," *J. Am. Heart Assoc.*, vol. 5, e002606, 2016.

[6] M.T. Brown, J. Bussell, S. Dutta, K. Davis, S. Strong, S. Mathew, "Medication adherence: Truth and consequences," *Am. J. Med. Sci.*, vol. 351, no. 4, pp. 387-399, 2016, doi: 10.1016/j.amjms.2016.01.010.

[7] M.E. Wilder, P. Kulie, C. Jensen, P. Levett, J. Blanchard, L.W. Dominguez, M. Portela, A. Srivastava, Y. Li, and M.L. McCarthy, "The impact of social determinants of health on medication adherence: A systematic review and meta-analysis," *J. Gen. Intern. Med.*, vol. 36, no. 5, pp. 1359-1370, May 2021, doi: 10.1007/s11606-020-06447-0. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33515188.

[8] P. Kardas, P. Lewek, and M. Matyjaszczyk, "Determinants of patient adherence: A review of systematic reviews," *Front. Pharmacol.*, vol. 4, p. 91, Jul. 2013, doi: 10.3389/fphar.2013.00091. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/23898295.

[9] M. E. Wilder, P. Kulie, C. Jensen, P. Levett, J. Blanchard, L. W. Dominguez, M. Portela, A. Srivastava, Y. Li, and M. L. McCarthy, "The impact of social determinants of health on medication adherence: A systematic review and meta-analysis," *Journal of General Internal Medicine*, vol. 36, no. 5, pp. 1359–1370, May 2021, doi: 10.1007/s11606-020-06447-0.

[10] W. Kanyongo, A. E. Ezugwu, "Feature selection and importance of predictors of non-communicable diseases medication adherence from machine learning research perspectives," Informatics in Medicine Unlocked, vol. 38, 2023, Art. no. 101232, https://doi.org/10.1016/j.imu.2023.101232.

[11]Chawa MS, Yeh HH, Gautam M, Thakrar A, Akinyemi EO, Ahmedani BK. The Impact of Socioeconomic Status, Race/Ethnicity, and Patient Perceptions on Medication Adherence in Depression Treatment. Prim Care Companion CNS Disord. 2020 Dec 10;22(6):20m02625. doi: 10.4088/PCC.20m02625. PMID: 33306887.

[12] H. Habehh and S. Gohel, "Machine learning in healthcare," *Curr. Genomics*, vol. 22, no. 4, pp. 291-300, Dec. 2021, doi: 10.2174/1389202922666210705124359.

[13] P.H. Luckett *et al.*, "Biomarker clustering in autosomal dominant Alzheimer's disease," *Alzheimers Dement.*, vol. 19, no. 1, pp. 274-284, Jan. 2023, doi: 10.1002/alz.12661.

[14] Y. Wang, Y. Zhao, T.M. Therneau, E.J. Atkinson, A.P. Tafti, N. Zhang, S. Amin, A.H. Limper, S. Khosla, and H. Liu, "Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records," *J. Biomed. Inform.*, vol. 102, p. 103364, Feb. 2020, doi: 10.1016/j.jbi.2019.103364.

[15] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2014.

[16] N. Negi and G. Chawla, "Clustering algorithms in healthcare," in *Intelligent Healthcare*, S. Bhatia, A.K. Dubey, R. Chhikara, P. Chaudhary, and A. Kumar, Eds. Cham, Switzerland: Springer, 2021, EAI/Springer Innovations in Communication and Computing, doi: 10.1007/978-3-030-67051-1_13.

[17] E. Ahlqvist, P. Storm, A. Käräjämäki, M. Martinell, M. Dorkhan, A. Carlsson, P. Vikman, R.B. Prasad, D.M. Aly, P. Almgren, Y. Wessman, N. Shaat, P. Spégel, H. Mulder, E. Lindholm, O. Melander, O. Hansson, U. Malmqvist, Å. Lernmark, K. Lahti, T. Forsén, T. Tuomi, A.H. Rosengren, and L. Groop, "Novel subgroups of adult-onset diabetes and their association with outcomes: A data-driven cluster analysis of six variables," *Lancet Diabetes Endocrinol.*, vol. 6, no. 5, pp. 361–369, May 2018, doi: 10.1016/S2213-8587(18)30051-2.

[18] K. Mittal, G. Aggarwal, and P. Mahajan, "Performance study of K-nearest neighbor classifier and K-means clustering for predicting the diagnostic accuracy," *Int. J. Inf. Technol.*, vol. 11, pp. 535–540, 2019, doi: 10.1007/s41870-018-0233-x.

[19] H. Alashwal, M. El Halaby, J.J. Crouse, A. Abdalla, and A.A. Moustafa, "The application of unsupervised clustering methods to Alzheimer's disease," *Front. Comput. Neurosci.*, vol. 13, p. 31, May 2019, doi: 10.3389/fncom.2019.00031.

[20] M.L. Ljubicic, A. Madsen, A. Juul, K. Almstrup, and T.H. Johannsen, "The application of principal component analysis on clinical and biochemical parameters exemplified in children with congenital adrenal hyperplasia," *Front. Endocrinol. (Lausanne)*, vol. 12, p. 652888, Aug. 2021, doi: 10.3389/fendo.2021.652888.

[21] P. Richette, M. Vis, S. Ohrndorf, W. Tillett, J. Ramírez, M. Neuhold, M. van Speybroeck, E. Theander, W. Noel, M. Zimmermann, M. Shawi, A. Kollmeier, and A. Zabotti, "Identification of PsA

phenotypes with machine learning analytics using data from two phase III clinical trials of guselkumab in a bio-naïve population of patients with PsA," *RMD Open*, vol. 9, no. 1, p. e002934, Mar. 2023, doi: 10.1136/rmdopen-2022-002934.

[22] S.W.M. Eng, F.A. Aeschlimann, M. van Veenendaal, R.A. Berard, A.M. Rosenberg, Q. Morris, et al., "Patterns of joint involvement in juvenile idiopathic arthritis and prediction of disease course: A prospective study with multilayer non-negative matrix factorization," *PLoS Med.*, vol. 16, no. 2, p. e1002750, 2019, doi: 10.1371/journal.pmed.1002750.

[23] H. Ma, J. Ding, M. Liu, and Y. Liu, "Connections between various disorders: Combination pattern mining using Apriori algorithm based on diagnosis information from electronic medical records," *Biomed. Res. Int.*, vol. 2022, p. 2199317, May 13, 2022, doi: 10.1155/2022/2199317.

[24] A.M. Ikotun, A.E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci.*, vol. 622, pp. 178–210, 2023, doi: 10.1016/j.ins.2022.11.139.

[25] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: An overview," *WIREs Data Mining Knowl Discov*, vol. 2, no. 1, pp. 86–97, 2012, doi: 10.1002/widm.53.

[26] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1650–1654, Dec. 2002, doi: 10.1109/TPAMI.2002.1114856.

[27] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *Ann. Data Sci.*, vol. 2, pp. 165–193, 2015, doi: 10.1007/s40745-015-0040-1.

[28] L. Kaufman and P. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, New York, NY, USA: Wiley, 1990, doi: 10.2307/2532178.

[29] M.E. Celebi, H.A. Kingravi, and P.A. Vela, "A comparative study of efficient initialization methods for the k-means clustering algorithm," *Expert Syst. Appl.*, vol. 40, no. 1, pp. 200–210, 2013, doi: 10.1016/j.eswa.2012.07.021.

[30] J.A. Hartigan and M.A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *J. Roy. Stat. Soc. C Appl. Stat.*, vol. 28, no. 1, pp. 100–108, 1979, doi: 10.2307/2346830.

[31] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: An overview," *WIREs Data Mining Knowl. Discov.*, vol. 2, no. 1, pp. 86–97, 2012, doi: 10.1002/widm.53.

[32] J.H. Ward Jr., "Hierarchical grouping to optimize an objective function," *J. Amer. Stat. Assoc.*, vol. 58, pp. 236–244, 1963, doi: 10.1080/01621459.1963.10500845.

[33] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: An overview," *WIREs Data Mining Knowl. Discov.*, vol. 2, no. 1, pp. 86–97, 2012, doi: 10.1002/widm.53.

[34] T. Prell, J. Grosskreutz, S. Mendorf, O.W. Witte, and A. Kunze, "Data on adherence to medication in neurological patients using the German Stendal Adherence to Medication Score (SAMS)," *Data Brief*, vol. 23, p. 103855, 2019, doi: 10.1016/j.dib.2019.103855.

[35] F. Feldmann, H. M. Zipprich, O. W. Witte, and T. Prell, "Self-reported nonadherence predicts changes of medication after discharge from hospital in people with Parkinson's disease," *Parkinson's Dis.*, vol. 2020, p. 4315489, 2020, doi: 10.1155/2020/4315489.

[36] S. Mendorf, O. W. Witte, H. Zipprich, and T. Prell, "Association between nonmotor symptoms and nonadherence to medication in Parkinson's disease," *Front. Neurol.*, vol. 11, p. 551696, 2020, doi: 10.3389/fneur.2020.551696.

[37] H.-M. Gerland and T. Prell, "Association between the health locus of control and medication adherence: An observational, cross-sectional study in primary care," *Front. Med.*, vol. 8, p. 705202, 2021, doi: 10.3389/fmed.2021.705202.

[38] H. M. Zipprich, S. Mendorf, A. Schönenberg, et al., "The impact of poor medication knowledge on health-related quality of life in people with Parkinson's disease: A mediation analysis," *Qual. Life Res.*, vol. 31, pp. 1473–1482, 2022, doi: 10.1007/s11136-021-03024-8.

[39] O. Mukhtar, J. Weinman, and S. H. D. Jackson, "Intentional non-adherence to medications by older adults," *Drugs Aging*, vol. 31, pp. 149–157, 2014, doi: 10.1007/s40266-014-0153-9.