HEALTHCARE PATIENT SEGMENTATION USING HIERARCHICAL CLUSTERING IN SPSS

1. Background and Problem Statement

A community healthcare clinic in the U.S. was looking to improve preventive care delivery and reduce unnecessary readmissions. Despite collecting extensive patient data—including visit frequency, diagnosis types, insurance coverage, and medication adherence—the clinic had no structured understanding of patient subgroups. The goal was to perform patient segmentation using hierarchical clustering in SPSS to identify vulnerable groups and personalize care coordination strategies accordingly.

2. Objectives

- To segment patients using hierarchical clustering in SPSS based on behavioral and demographic health indicators
- To identify high-risk patients who require proactive care plans and follow-ups
- To assist clinic management in resource planning and targeted patient communication
- To visualize segment characteristics using SPSS dendrograms and descriptive statistics

3. Methodology

3.1 Dataset Description

- Source: Electronic Health Record (EHR) data from 1,200 patients collected over the past 18 months
- Variables used:
    - Age
    - Number of doctor visits
    - Presence of chronic conditions (binary-coded: diabetes, hypertension, asthma)
    - Insurance status (public/private/uninsured)
    - Missed appointment rate (%)
    - Medication adherence index

3.2 Data Preparation in SPSS

- Categorical variables (e.g., insurance type) converted to dummy variables
- Outliers and missing values addressed using mean substitution and range validation
- Standardized numerical variables using z-scores to remove unit bias

3.3 Hierarchical Clustering in SPSS

- Distance measure: Squared Euclidean
- Method: Ward's linkage for minimizing intra-cluster variance
- Scree plot and agglomeration schedule used to determine the optimal number of clusters (k=4)
- Dendrogram generated for visual confirmation of distinct clusters

4. Results

Cluster Profiles:

- **Cluster 1 – Elderly Chronic Patients (20%)**
  - High visit frequency, multiple chronic conditions, low medication adherence
  - Mostly publicly insured or uninsured
- **Cluster 2 – Young Adherent Patients (25%)**
  - Low visit frequency, high adherence, no chronic illness
  - Majority with private insurance
- **Cluster 3 – At-Risk Non-Adherent Adults (35%)**
  - Moderate age, high no-show rate, low adherence
  - Cost-conscious and insurance-variable
- **Cluster 4 – Infrequent Visitors with One Condition (20%)**
  - Younger to middle-aged, one manageable condition, medium adherence
  - Least engaged with preventive care

5. Interpretation and Insights

- Cluster 1 patients have the highest resource utilization and need better medication management

- Cluster 3 presents the greatest risk for preventable ER visits due to non-adherence
- Cluster 2 shows stable health behavior; suitable for digital self-management tools
- Cluster 4 can be nudged into more regular checkups with health reminders

6. Recommendations

- **Case managers** should prioritize outreach to Clusters 1 and 3
- Tailor follow-up call scripts and educational material based on cluster characteristics
- Design a **mobile reminder system** and SMS alerts for Clusters 3 and 4
- Expand preventive care campaigns for Cluster 2 to maintain their health status

7. Future Work

- Integrate psychological and lifestyle variables to enrich cluster profiles
- Explore supervised classification (e.g., decision trees) to predict future cluster movement
- Validate clusters against hospitalization and readmission records
- Re-assess clustering quarterly to monitor population dynamics

8. Stakeholder Relevance

Academic:

- Illustrates hierarchical clustering for healthcare segmentation using real-world variables
- Suitable for coursework in public health analytics, biostatistics, and clinical decision-making

Corporate (Healthcare Providers and Clinics):

- Enables patient-centric care coordination
- Supports targeted investment in telehealth, chronic disease management, and outreach programs