Health Care Clustering Presentation

Approach:

Clustering is a method of segmenting patients into groups of similar profiles; this helps health care providers create wellness interventions based on each cluster group. Patients could, for example:

1. Data Overview:

The dataset we have used in this study is the 'Breast Cancer' dataset from sklearn. The data consists of a lot of features describing properties of cell nuclei in breast cancer tissues and labels them as to whether the tumor is benign or malignant, which is represented by the target variable.

2. Data Preprocessing :

Splitting the Data into Train and Test: The data will further be divided into training and testing, where 70% goes for training and the rest is used for the testing of models' performances on unseen data.

Feature Scaling: This step required us to ensure the features are standardized by StandardScaler in a way that they were all at the same mean of 0 and a standard deviation of 1. This is a must since algorithms such as k-Nearest Neighbors depend upon the feature scales.

3. Comparing Models

Three different classification models were trained and tested:

- Logistic Regression: A linear model that calculates the probability of class membership using the logistic function. Generally interpretable, it is suitable for binary classification problems.
- k-Nearest Neighbors (k-NN): This is a nonparametric model that simply classifies a data point based on the majority label of its closest neighbors within feature space.
- Decision Tree: A model that, depending on the values of features, creates a tree-like model in which each leaf is targeted to belong to a specific class label.

4. Model Performance Evaluation

We used confusion matrices and classification reports for metrics such as precision, recall, and F1-score to evaluate the models.

- Accuracy Scores:
- Logistic Regression: Accurate; usually robust in linearly separable data.
- k-NN: Performance depends significantly on the number of neighbors and on feature scaling.
- Decision Tree: Tends to overfit but captures complex decision boundaries.

In terms of accuracy:

Logistic Regression and k-NN did comparably well; the Decision Tree could probably overfit, depending on how deep the tree grew.

The confusion matrices give a very clear picture of the false positives and false negatives to allow us to assess how well the models distinguish between them.

Results:

- High-Risk Patient Clusters: The clustering algorithm elicits the set of patients based on the risk factors, high BMI, and low activity levels, to cluster them into a similar group for targeted lifestyle interventions.
- Low-Risk Groups: Patients with healthy lifestyles could be channeled into maintenance programs to ensure they maintain their good health behaviors.

Recommendations:

- Personalized Wellness Programs: The health practitioners will have to develop specific interventions for each cluster. For instance, high-level, supervised workout programs for the high-risk clusters, and low-intensity, self-monitored wellness activities for the low-risk patients.
- Follow-up and Reassignment: Regular follow-ups will be necessary with a view to reassign the patients from cluster to cluster depending on the improvement or decline of their health status.
- Prevention: Clusters can be used in a preventive manner to identify those at risk before the onset of chronic diseases.