Machine Learning Clustering Final Project

By: David Wilczynski

1. Main Objective

The main objective of this analysis is to utilize three different clustering techniques for unsupervised machine learning. Clustering in this example will be used to group individuals into meaningful risk segments for insurance based on lifestyle, demographic, and health factors as covariates. The clusters in a clinical setting could be used to help insurance providers identify high-risk populations in order to develop personalized plans and improve pricing strategies. Like previously mentioned, this analysis will be focused on clustering, specifically three different clustering techniques from Sklearn library in Python.

1. Description of the Data

The dataset for this analysis comes from the UC Irvine Machine Learning Repository entitled Medical Cost Personal Dataset which contains 1338 participants along with various demographic variables including age, sex, BMI, number of dependents, smoking status, region of the country, and final insurance bill (which is left out of the analysis for clustering purposes, but will be used to cross check the results). The goal is to identify patterns in the dataset by clustering individuals into risk groups, so that the results could potentially be used to predict new participants by seeing where they fall into the clusters.

In terms of Data exploration and preprocessing, the categorical variables of sex, smoking status, and region of the country were handled using one-hot encoding. The numeric variables on the other hand including age, BMI, and number of dependents were standardized using StandardScaler. As previously mentioned, the charges variable which codes the final insurance total for each patient was not included in analysis but utilized in order to cross check the clusters to see if the objective of this analysis was met.

For the analysis, three different clustering techniques were run and compared based on the results to see which is the best fit for the data. The three clustering techniques utilized were K-Means Clustering, Hierarchical Clustering, and DBSCAN.

1. Clustering Models

First, K-Means Clustering was used as the first unsupervised Machine Learning technique. The clustering was run using values K=2 through K=10 with the Elbow Method in order to see what the appropriate value of K should be for this data. Ultimately, K=4 was chosen as it produced well separated interpretable clusters which could also be visualized using Python. The results of the Elbow Method and the PCA projections for the K-Means clusters can be seen below.

A graph with a line

AI-generated content may be incorrect.

A diagram of different colored dots

AI-generated content may be incorrect.

Secondly, a Hierarchical Agglomerative Clustering method was used utilizing a Ward linkage technique. This method also resulted in 4 clusters which can be seen below in the plot.

A graph with colored lines

AI-generated content may be incorrect.

Finally for this analysis, the DBSCAN method was utilized in order to cluster the data for insurance charge purposes.

A diagram of multiple colored dots

AI-generated content may be incorrect.

1. Comparison and Key Findings

After evaluating the three different clustering techniques utilized in this analysis, K-Means clustering would be the recommended method as it resulted in clear cut clusters, and the final number of clusters could be adjusted based on the needs of the insurance company or if more data was included in the analysis at a later time. After the analysis was finished, the 4 different clusters resulting from the K-Means clustering can be summarized as seen below and cross checked against the insurance charge variable which was not included in the analysis.

A number of numbers on a white background

AI-generated content may be incorrect.

A chart of distribution of charges

AI-generated content may be incorrect.

In terms of next steps for the analysis, it would be important to include more data within the models like medical history and exercise as this clustering example was based on only a few variables. Since the insurance charges were ultimately a variable within this dataset, it could also be beneficial to test these models against Supervised Machine Learning Techniques.