D212: Data Mining II

Performance Assessment 1

Part A: Research Question

1. One question that could be asked by an organization for this data would be to see how income affects health factors like Vitamin D levels as well as charges that are accumulated at the hospital like the additional charges or even days stayed in the hospital.
2. One of the goals with this analysis is to figure out how hierarchical clustering could help us understand the relationships between Income and health factors as well as if income has any affect on the charges that are accumulated by patients with higher incomes.

Part B: Technique Justification

1. I chose to use a hierarchical clustering technique because it’s a much better way to figure out how many clusters might be present in a dataset. Using a dendrogram in a hierarchical analysis, you can get a good idea of how many clusters are related and how many there are under certain circumstances. Expected outcomes of hierarchical clustering is to figure out distances between points and to see which factors tend to show greater relations which other elements within the analyzed dataset.
2. One assumption that needs to be made within a hierarchical clustering analysis is that there is a measure of dissimilarity between datapoints as well as a linkage between the datapoints that can be measured using a form of distance measurement.
3. I used python for my analysis as it’s the language I’m most familiar with. I used numpy for math operations, pandas for database operations, matplotlib for more math operations and graphing, seaborn for more graphing options, and sklearn and scipy for the stats and machine learning libraries needed for this analysis to work.

Part C: Data Preparation

1. The first part of any data analysis is checking for null values and outliers in order to strip away anything that could mess with the analysis. For hierarchical clustering, null values can cause problems so I made sure to check for null values and deleted the rows that contained them which allowed the analysis to continue along smoothly.
2. For this analysis I used Income, VitD\_levels, Doc\_visits, Initial\_days, TotalCharge, and Additional\_charges in order to see what factors income could affect. All of these variables are continuous since categorical values can’t really be normalized, which is needed for my analysis.
3. In order to prepare my data for analysis, I first had to isolate the variables I wanted to use in my analysis. I was able to create a new datafile with the columns I wanted and then normalized all the variables because if it wasn’t normalized, 1 variable could drastically swing the results. The next step was to create a dendrogram to see what kind of relationships exist between the variables already. I also created a clustered scatter plot, and a heatmap in order to see some quick relationships to help improve the analysis. The figures are attached. I then started exploring by comparing different factors of the data and seeing the cluster relationships between them. The color split was useful in identifying which factor is which when all the points are clustered on a diagram. Attached is all the code and diagrams used in this analysis.
4. Attached is a copy of the cleaned dataset.

Part D: Analysis

1. For my analysis I did a hierarchical clustering of various columns of data. The main focus was to see if Income had any affects on charges at the hospital or health factors. I was able to compare these factors in a few charts that are attached. I first normalized the data in order to prevent one factor from being weighted to heavily or not enough. To normalize I just used the simple normalization equation of a standard score which is where x is the observed value, µ is the mean of the sample, and σ is the standard deviation of the sample. This helped normalize our variable with respect to each other variable and then found the Euclidean distance between the points using the equation where q and p are the Euclidean vectors and n is the number of spaces that the measurement is taking place over.
2. See attached for the code

Part E: Summary and Implications

1. The accuracy of this analysis can be assessed by seeing how clustered our actual clustering is. Because I used wards method for my hierarchical clustering which uses a summed square calculation in order to depict the dendrogram which can then be translated into scatter plots for easier understanding. When it comes to accuracy, the primary way of finding the accuracy is simply by observing what graphs are presented. Searching the cluster can give a good idea on how related the factors being compared are and how similar or different they are in relation to other factors.
2. One of the results that I have found from this analysis is that Income seems to be very unrelated to multiple factors. After a basic heatmap was created, I tried plotting the clusters of income with things like Vitamin D levels and additional charges, but there seemed to be very little relationship between them. I also tried testing other factors to see if a relationship was formed, and if I compared doctor visits with initial days in the hospital, I found 2 unique clusters while everything else with income was just 1 cluster, showing really no relationship between income and other factors.
3. One limitation of this analysis is the dendrogram. It can be a difficult thing to understand and effectively use for an analysis. One of the biggest issues with hierarchical clustering is the lack of knowing how many clusters should exist. K-means clustering allows you to specify how many clusters to use in an analysis while hierarchical you start with each sample as its own cluster and then keep joining the most similar samples.
4. The recommended course of action for my organization would be to not worry about income when it comes to health outcomes. It appears that there isn’t much affect on prices or health outcomes in the few variables that I tested. This does not mean Income has no affect on anything at the organization, it’s just that my clustering analysis came up short and I did not find a relationship between income and the other variables.

Part E: Video

Panopto video: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=8ebb3be7-6b46-4160-a2b5-ae290183ea37>

Sources:

<https://www.globaltechcouncil.org/clustering/k-means-clustering-vs-hierarchical-clustering/> K-means clustering vs hierarchical clustering, Global Tech Council

<https://www.stat.purdue.edu/bigtap/online/docs/Hierarchical_Clustering_Complete.html> Hierarchical Clustering, Purdue University

<https://stats.stackexchange.com/questions/195446/choosing-the-right-linkage-method-for-hierarchical-clustering> Choosing the right linkage method for hierarchical clustering, ttnphns

<https://www.analyticsvidhya.com/blog/2021/08/hierarchical-clustering-algorithm-python/> Hierarchical Clustering Algorithm Python, Hardikkumar Dhaduk

<https://www.youtube.com/watch?v=7xHsRkOdVwo> StatQuest: Hierarchical Clustering, StatQuest with Josh Starmer

<https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/> Hierarchical Clustering with Python and Scikit-Learn, Usman Malik

<https://www.analyticsvidhya.com/blog/2019/05/beginners-guide-hierarchical-clustering/> A Beginner’s Guide to Hierarchical Clustering and how to Perform it in Python, Pulkit Sharma