Gaby Masak

D603 – Machine Learning

Task 2: Clustering Technologies

11/23/2024

Explanation: Code for Data Production Pipeline

**Requirement A: Gitlab Subgroup and Project**

GitLab URL: <https://gitlab.com/wgu-gitlab-environment/student-repos/gmasak/d603-machine-learning/-/commits/working_branch>

Screenshot of Repository Branch History:

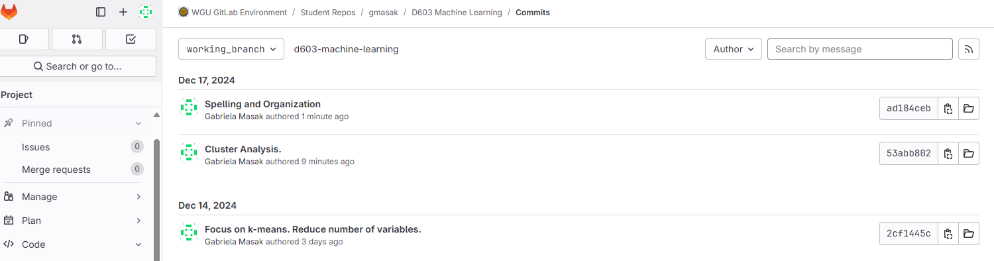
****

Figure 1: Screenshot of Repository Branch History

**Requirement B: Purpose of Data Mining Report**

Hospitals, like many businesses and organizations, benefit from understanding the characteristics of their patients and clientele. By collecting and leveraging larges sets of data on treatment and finances, it is possible to apply k-means clustering methods to answer questions surrounding groupings of which financial characteristics are most relevant in order to derive more targeted business solutions and treatments. The goal of using this data to apply more targeted treatment can improve patient outcomes and reduce the long-term cost of care.

**Requirement C: Reasons for Clustering Method**

K-means clustering is an unsupervised learning method used to partition complex datasets into distinct clusters based on their similarities. This method aims to minimize the variance within each cluster while maximizing the variance between clusters. The process begins by selecting a specified number of cluster centroids (k) and assigning each data point to the nearest centroid. The centroids are then recalculated based on the mean of the assigned points, and the assignment process is repeated until convergence is achieved (Glen, 2016). Expected outcomes include the identification of natural groupings within the data, which can reveal underlying patterns and relationships that may not be immediately apparent, assuming that the data is appropriately preprocessed and the chosen distance metric and linkage method are suitable for the dataset (Pai, 2021).

Python packages:

1. **pandas**:

* **Purpose**: Data manipulation and analysis.
  + pandas provides data structures like dataframes for data handling and processing.

1. **scikit-learn**:

* **Purpose**: Machine learning and data mining.
  + scikit-learn offers a various tools, including StandardScaler and OneHotEncoder, for model building and evaluation.

1. **matplotlib**:

* **Purpose**: Data visualization.
  + matplotlib is used to create visualizations such as scatterplots, among other plots and graphs.

1. **seaborn**:

* **Purpose**: Statistical data visualization.
  + seaborn builds on matplotlib and provides an interface for drawing statistical graphics, such as heatmaps.

1. **Numpy**:

* **Purpose**: Numerical computing.
  + numpy provides mathematical functions to operate on arrays.

1. **SciPy**:

* **Purpose**: Statistical modules for statistics, optimization, integration, linear algebra.
  + scipy provides methods cluster the data via fcluster, linkage, and dendrogram.

**Requirement D: Data Preparation**

As discussed above, several statistical clustering methods from SciPy were essential in the development and implementation of k-means clustering models on the health data set. These require continuous variables as input, reducing the number of variables to 7. Further reduction was used to focus on the financial continuous variables, as seen in Table 1 below.

Table 1: Variable Classification

|  |  |
| --- | --- |
| **Variable** | **Classification (Continuous/Categorical)** |
| Income | Continuous |
| Initial\_days | Continuous |
| TotalCharges | Continuous |
| Additional\_charges | Continuous |

Several steps were taken to prepare the data for analysis. First, any duplicate rows are removed, identified by the code segment in Figure 2 below.



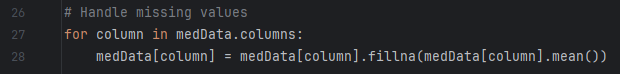
Figure 2: Screenshot of Code Removing Duplicate Rows

Next, variable selection reduced the 49 independent variables into 4 by removing any variables that were not continuous or relevant.



Figure 3: Screenshot of Code Removing Redundant Variables

Subsequently, any missing values are filled. For missing values in numerical columns, the mean of the column is used, as seen in Figure 4 below.

A screen shot of a computer code

Description automatically generated

Figure 4: Screenshot of Code Replacing Missing Values

Next, the data was standardized. A copy of the cleaned dataset can be found in the attached cleaned\_data.csv file.

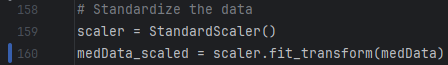


Figure 5: Screenshot of Standardization

**Requirement E: Data Analysis**

After the data is preprocessed, under K-means clustering, the optimized value can be expressed as the number of clusters that maximizes the silhouette score. The silhouette score measures how similar an object is to its own cluster compared to other clusters. This value ranges from -1 to 1, with higher values indicating better-defined clusters. In Figure 6, the optimal number of clusters is represented by 2 clusters, as this configuration yields the highest silhouette score.

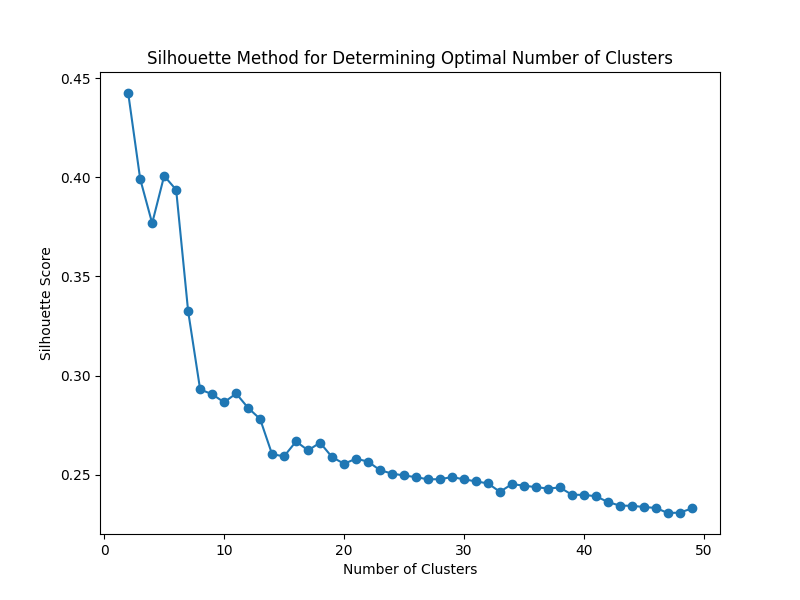


Figure 6: Silhouette Scores of Data

**Requirement F: Summary**

As discussed previously, k-means clustering was applied to the dataset. The resulting visualizations of the clusters allows for analysis of cluster quality. After applying principal component analysis, it is possible to visualize the clusters, seen in Figure 7 below. Two clusters are defined by navy and teal colors. The majority of the data points are neatly clustered together, but both clusters have small satellite data points below the main clusters, as well as a few points scattered in the area between.

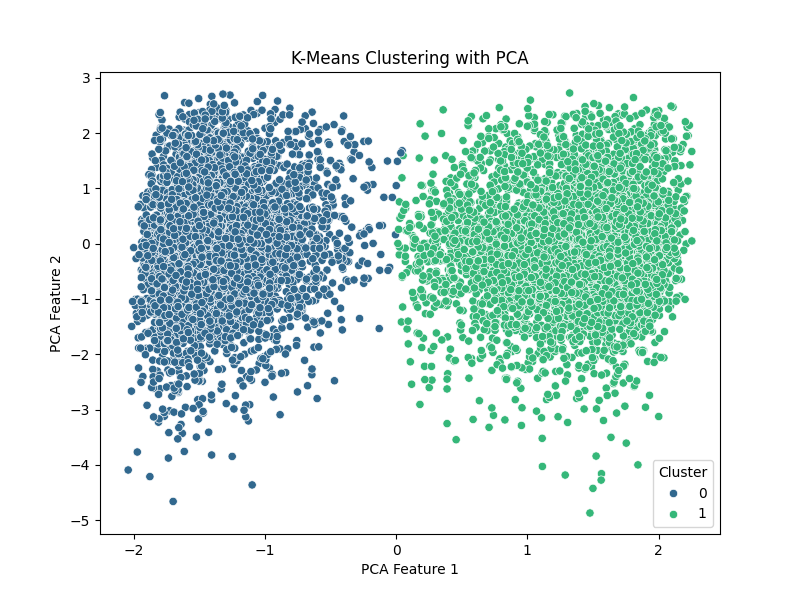


Figure 7: K-Means Clusters

Further analysis can be taken by counting the number of data points in each cluster and comparing as understanding the sizes of the clusters provides insight into the overall structure and composition of the dataset. Disparity in the sizes of clusters can point towards common trends in larger clusters or outliers and minority subgroups in smaller clusters. There are 5002 data points in Cluster 1 and 4998 data points in Cluster 2 as seen in Figure 8 below.

A screenshot of a computer

Description automatically generated

Figure 8: Cluster Analysis

The quality of the clusters can be assessed by using the Davies-Bouldin Index (DBI). DBI evaluates clusters for compactness and separation. Lower values are optimal, representing well-separated and compact clusters. The DBI of the clusters in Figure 7 is calculated to be 0.9759 which indicates that the clusters are well-defined and distinct from each other. The compactness of the clusters suggests that the data points within each cluster are similar to one another, while the separation indicates that the clusters are distinct and not overlapping. This level of cluster quality is crucial for ensuring that the clustering results are meaningful and can be used to derive actionable insights.

The clusters can further be dissected by interpreting the characteristics represented by the clusters. Cluster 1 consists of the following descriptive data: the average income in this cluster is approximately $40,737, with a standard deviation of $28,667, indicating a wide range of income levels among the patients. The minimum income is $154, while the maximum income is $204,542. The average number of initial days is about 9.24, with a standard deviation of 6.16 days. The total charges for patients in this cluster average around $3,248, with a standard deviation of $607. The additional charges have a mean of $12,895 and a standard deviation of $6,582. The wide range in income and additional charges suggests significant variability in the financial aspects of the patients in this cluster. The median values for income, initial days, total charges, and additional charges are $33,984, 7.9 days, $3,179, and $11,535, respectively, indicating that half of the patients have values below these medians.

For Cluster 2, the average income in this cluster is approximately $40,244, with a standard deviation of $28,375, similar to Cluster 0. The minimum income is $301, while the maximum income is $207,249. The average number of initial days is significantly higher at about 59.69 days, with a standard deviation of 8.58 days, indicating that patients in this cluster tend to stay longer initially. The total charges for patients in this cluster average around $7,378, with a standard deviation of $782. The additional charges have a mean of $12,974 and a standard deviation of $6,503. The median values for income, initial days, total charges, and additional charges are $33,628, 61.17 days, $7,461, and $11,643, respectively. Compared to Cluster 1, the higher average and median values for initial days and total charges suggest that patients in this cluster require more extensive and costly medical care.

The implications of this cluster analysis are significant for understanding the dataset's structure and making informed business and healthcare decisions based on the clustering results. For instance, Cluster 1, with an average income of approximately $40,737 and shorter initial stays (mean of 9.24 days), suggests that these patients might have less severe medical conditions or require less intensive care. In contrast, Cluster 2, with a similar average income of $40,244 but significantly longer initial stays (mean of 59.69 days) and higher total charges, indicates more severe conditions or more intensive medical care. These insights can help healthcare providers tailor their services and allocate resources more effectively, as income and additional charges are similar in both clusters, indicating that they are less likely to be the primary factors influencing the length of stay and total charges. Instead, initial number of days hospitalized appears to be the driving factor behind increased total charges. By understanding these distinctions, healthcare providers can develop targeted interventions and allocate resources more efficiently to address the specific needs of each patient group, ultimately improving patient outcomes and optimizing healthcare delivery.

The discrepancies observed in the k-means clustering highlight the importance of carefully selecting the variables and clustering method used and considering the potential for misclassification. This careful selection should also be acknowledged in terms of limitations for the analysis. While this analysis only focused solely on financial variables, additional demographic variables could be included to provide a more comprehensive or generalized analysis on patients and their care. Other clustering methods such as hierarchical clustering would be required due to the addition of non-continuous data. Additional clustering analysis may also enhance the robustness of the analysis and yield new insights.

In summary, this analysis underscores the need for a thorough evaluation of clustering results and the consideration of multiple clustering techniques to ensure robust and reliable insights. Ultimately, the choice of clustering method and parameters should be guided by the specific goals of the analysis and the characteristics of the dataset. By leveraging k-means clustering, the hospital can gain deeper insights into financial data, identifying natural groupings that may not be immediately obvious. This can lead to more personalized financial decisions, as clusters may reveal specific financial profiles. Additionally, understanding these clusters can help in resource allocation, ensuring that the hospital is better prepared to meet the needs of different patient groups as local communities and hospital services grow and develop.

**Requirement G: Panopto Video Link**

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=8f71382e-2196-4b18-a959-b24901291d1f>

**Sources Cited**

**Code:**

* + 1. *fcluster — SciPy v1.14.0 Manual*. (n.d.). Docs.scipy.org. <https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.fcluster.html>
    2. HeroVired Editorial Team. (2024, November 25). *Explore Hierarchical Clustering in Machine Learning*. Hero Vired; HeroVired. <https://herovired.com/learning-hub/blogs/hierarchical-clustering-in-machine-learning/>
    3. Jacobs. (2023, August 20). *Clustering on Principal Component Analysis*. Charles

Holbert. <https://www.cfholbert.com/blog/cluster-pca/>

* + 1. Khyati Mahendru. (2019, June 17). *How to determine the optimal K for K-Means?* Medium; Analytics Vidhya. <https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb>
    2. *scipy.cluster.hierarchy.linkage — SciPy v1.8.0 Manual. (n.d.). Docs.scipy.org.* [*https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html*](https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html)

**Report:**

1. Glen, Stephanie. (2016, November 15). *Hierarchical Clustering / Dendrogram: Simple*

*Definition, Examples*. Statistics How To. <https://www.statisticshowto.com/hierarchical-clustering/>

1. Pai, P. (2021, May 10). *Hierarchical clustering explained*. Medium.

https://towardsdatascience.com/hierarchical-clustering-explained-e59b13846da8