**Task 2: Clustering Techniques**

Joseph Emmanuel Cayetano

Western Governors University

D603: Machine Learning Task 2

January 31, 2025

1. **Create your subgroup and project in GitLab**

**GitLab Repository URL:** https://gitlab.com/wgu-gitlab-environment/student-repos/jcayet5/d603-machine-learning/-/tree/task2branch?ref\_type=heads

**Repository Branch History Screenshot**

A screenshot of a computer

Description automatically generated

**B1. Propose one question relevant to a real-world organizational situation that you will answer using the K-means clustering technique**

One question that is relevant to a real-world organizational situation that I will answer using the K-means clustering technique is: What distinct patient groups, based on demographic, health conditions and hospitalization details, can help the hospital create targeted strategies to reduce readmissions and lower hospital costs?

**B2. Define one goal of the data analysis**

The goal of the data analysis is to identify patient groups with similar demographics, medical conditions, and hospitalization details linked to higher readmission rates, helping the hospital create targeted strategies to reduce readmissions and lower hospital costs.

This goal is reasonable because it supports the hospital’s need to reduce readmissions and lower costs, as outlined in the scenario. The goal is also represented in the medical dataset, as the dataset includes relevant variables such as patient demographics, medical conditions, and hospitalization details, which are needed for K-means clustering analysis.

**C1. Explain how the clustering technique you chose analyzes the selected dataset. Include expected outcomes**

The K-means clustering technique analyzes the medical dataset in multiple steps. The first step involves selecting key variables such as patient demographics, medical conditions, and hospitalizations details. These variables will be used as input variables for clustering. The second step involves normalizing the data. Since some variables in the dataset have different scales like income and age, the data will be normalized to ensure they do not dominate the clustering process. Categorical variables must be encoded numerically, as K-means clustering requires the use of continuous variables only. The third step involves cluster formation. The K-means algorithm starts by “randomly placing cluster centers (centroids), then assigns each data point to the closest centroid” (Banoula, 2024, par. 78). The centroids are recalculated iteratively until the clusters stabilize. The fourth step involves cluster evaluation. The clusters are analyzed to identify characteristics specific to each group.

Using the K-means clustering technique on the medical dataset yields multiple outcomes. The first outcome involves getting distinct patient clusters. The clusters will represent groups of patients with similar demographics, medical conditions, and hospitalization details. For example, one cluster could include patients with high readmission risk due to multiple chronic conditions, while another might group patients with short hospital stays and minimal complications. The second outcome involves getting specific characteristics from each cluster. This helps the hospital focus on preventive care and create tailored programs for high-risk patients. The third outcome involves cost-efficiency improvements. By understanding the characteristics of each cluster, the hospital can take targeted actions to reduce penalties, improve care, and control costs effectively.

**C2. Summarize one assumption of the clustering technique**

One assumption of K-means clustering is that clusters are spherical, which means the data points in each cluster are distributed evenly around the centroid in all directions. The centroid is the average of all data points in a cluster. This means the algorithm works best with clusters that are similar in size and shape, not elongated or irregular.

**C3. List the packages or libraries you have chosen for Python, and justify how each item on the list supports the analysis**

There are several Python libraries I used for the K-means clustering analysis. First is the pandas library, and it was used to load and manipulate the medical dataset. The get\_dummies() function from pandas converted the categorical variables into numerical format for K-means clustering. Another key pandas function is mean(), as it calculated the average values of all numerical features within each cluster. Second is the NumPy library, and it converted the medical dataset from a pandas DataFrame into a NumPy array, which is required for K-means clustering. NumPy was also used in computing cluster centroids and standardizing data. Third is the sklearn library, and it contains the K-means algorithm and StandardScaler() function. The KMeans() function from sklearn implemented K-means clustering to group patients based on demographic and health-related attributes. The StandardScaler() function standardized the continuous variables in the medical dataset to ensure fair contribution of features to distance calculations. Fourth is the matplotlib library, and it was used to create the elbow method plot to determine the optimal number of clusters. Matplotlib also helped visualize cluster distribution using a scatter plot. Fifth is the seaborn library, and it enhanced the aesthetics of the scatter plot for cluster visualization by adding color palettes and a legend.

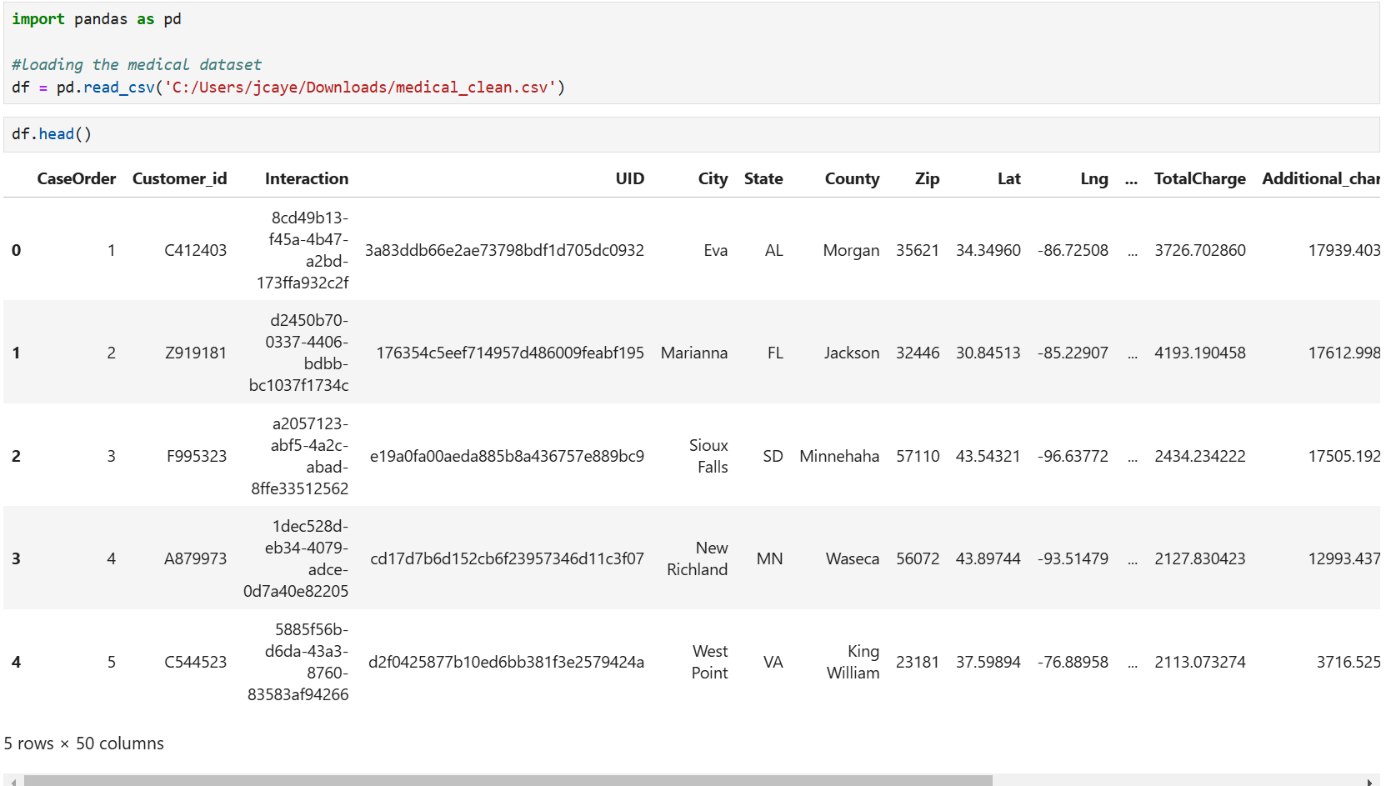
**D1. Describe one data preprocessing goal relevant to the clustering technique from part B1**

One data preprocessing goal for K-means clustering is normalizing or scaling the selected variables to ensure all variables contribute equally to the clustering process. K-means clustering uses Euclidean distance to measure the similarity between data points. Without normalization or standardization, features with larger ranges like income could dominate the clustering process, overshadowing features with smaller ranges like age. This is why applying normalization or standardization is important, as it puts all features on the same scale, ensuring fair and meaningful clusters.

**D2. Identify the initial dataset variables you will use to perform the analysis for the clustering question from part B1, and label each as continuous or categorical**

The variables that I will use to perform the analysis for the K-means clustering question from part B1 are: Age (continuous), Income (continuous), Gender (categorical), Marital (categorical), Area (categorical), HighBlood (categorical), Stroke (categorical), Overweight (categorical), Arthritis (categorical), Diabetes (categorical), Initial\_admin (categorical), Initial\_days (continuous), Complication\_risk (categorical), TotalCharge (continuous), and Additional\_charges (continuous).

**D3. Explain each of the steps used to prepare the data for the analysis. Identify the code segment for each step**



I imported and loaded the medical dataset using the read\_csv() function from pandas. The output from the head() function confirms that the dataset is loaded successfully. The dataset has a total of 50 columns.

A screenshot of a computer program

Description automatically generated

These are the variables I selected for analysis: Age, Income, Gender, Marital, Area, HighBlood, Stroke, Overweight, Arthritis, Diabetes, Initial\_admin, Initial\_days, Complication\_risk, TotalCharge, and Additional\_charges. I selected a total of 15 columns. The dataset is filtered to include only these selected variables, removing the rest. The output from the info() function confirms the selected variables exist and shows their data types and non-null value counts.

A screenshot of a computer screen

Description automatically generated

I checked for missing values by using both isnull() and sum() functions together. The dataset has no missing values, which means we can proceed directly to encoding categorical variables.

A screenshot of a computer program

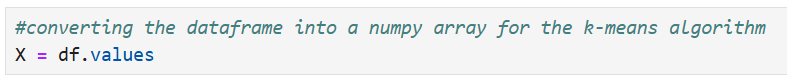
Description automatically generated

I applied one-hot encoding to convert categorical variables into numerical format, as K-means requires numerical data. After encoding, I used the rename() function from pandas to rename the column names for better readability. Next, I converted the Boolean values in the encoded columns to integers, so they appear as 0 or 1.

A screenshot of a computer

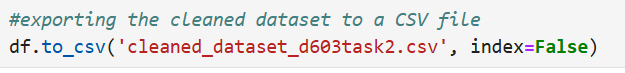
Description automatically generated

I used the StandardScaler() function from the sklearn library to standardize the original continuous columns by transforming them to have a mean of 0 and a standard deviation of 1. This is to ensure all variables contribute equally to the clustering process. The output from the head() function verifies that the original continuous variables have been transformed successfully.



In this code, I converted the DataFrame into a NumPy array, which is the format required by the K-means algorithm.

**D4. Provide a copy of the cleaned dataset**



The file cleaned\_dataset\_d603task2.csv is the cleaned medical dataset prepared for K-means clustering analysis and I included it in my submission.

**E1. Determine the optimal number of clusters in the dataset, and describe the method used to determine this number**

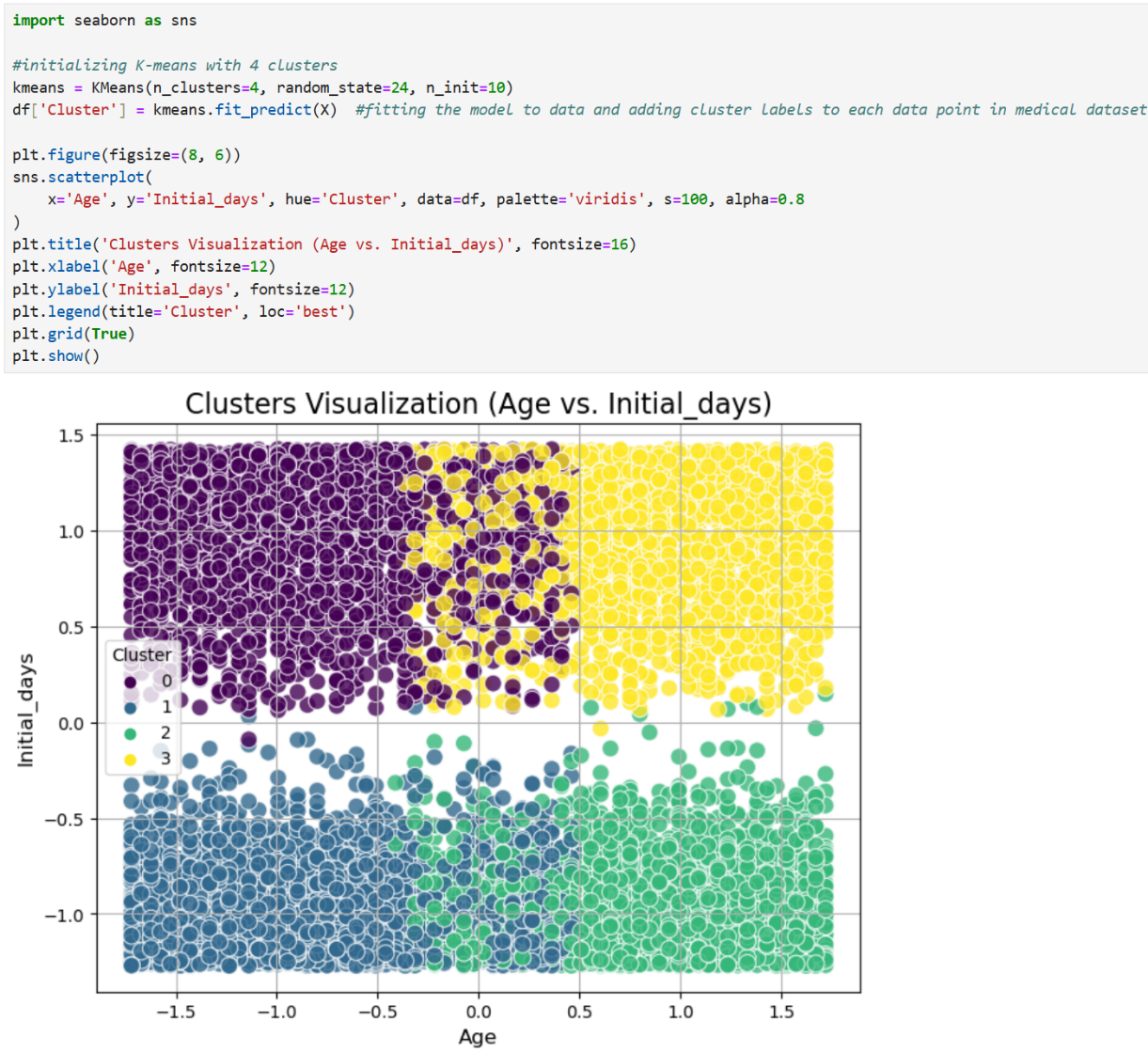
A graph and chart with text

Description automatically generated

In this code, I used the Elbow Method to identify the optimal number of clusters in the medical dataset. I started by converting the DataFrame into a NumPy array for the K-means algorithm. Then, I created an empty list to store inertia values for each k (number of clusters) and set the range of k to test from 1 to 10. Next, I created a for loop that iterates through each k from 1 to 10. Within the loop, a K-means model is created for each k, fitted to the data, and its inertia is added to the list. After the loop has finished, I plotted the results of the Elbow Method.

From the plot, the optimal number of clusters is 4. I used the Elbow Method to reach this decision. In this method, the "elbow" point represents the optimal number of clusters, where the decrease in inertia slows down significantly. Adding more clusters beyond this point provides minimal improvement in reducing inertia, making it inefficient. The reason why I selected 4 as the optimal number of clusters is because the inertia decreases steeply from k = 1 to k = 4, but the reduction becomes less significant after k = 4. This is essentially the "elbow" point in the plot. This suggests that 4 clusters provide a good balance between simplicity and accuracy.

**F1. Visualize the clusters and explain the quality of the clusters created. Include a screenshot of the cluster visualizations**



In this code, I performed K-means clustering on the medical dataset and visualized the clusters using two variables, which are Age and Initial\_days. I started by initializing K-means with 4 clusters. Then, I fit the K-means algorithm to the data and assigned each data point to one of the 4 clusters. The clusters labels (0, 1, 2, or 3) are stored in a new column called Cluster in the DataFrame. Next, I created the scatterplot that uses the Age and Initial\_days variables to visualize the clusters in two dimensions.

The clusters in the scatterplot are visually distinct and well-separated. There is a minimal overlap between clusters, which indicates that K-means has successfully grouped data points with similar characteristics. Each cluster is very compact, as the data points within a cluster are close to their respective centroids. The clusters are also fairly balanced, with no single cluster dominating excessively. Each cluster represents unique characteristics based on Age and Initial days. Cluster 0 (purple) represents younger patients with longer hospital stays, likely due to complex medical issues. Cluster 1 (blue) represents younger patients with shorter hospital stays, likely healthier or needing less intensive care. Cluster 2 (green) represents older patients with shorter hospital stays, likely due to less severe conditions or efficient treatment. Cluster 3 (yellow) represents older patients with longer hospital stays, likely due to severe or chronic health conditions needing extended care.

A screenshot of a computer

Description automatically generated

This is the cluster summary for 4 clusters created using the K-means algorithm. Each cluster represents a group of patients with unique characteristics based on demographics, hospital stays, and health conditions. In the cluster summary table, each row represents a cluster, and the columns show the mean or proportion values of features for the data points in that cluster. Continuous features show mean values, while categorical features show proportion values. Cluster 0 consists of patients with below-average age (-0.76) and income (-0.02), longer hospital stays (0.95), lower additional charges (-0.68), moderate rates of high blood pressure (0.30) and stroke (0.19), and average emergency admission rates (0.49). Cluster 1 consists of patients with below-average age (-0.76), above-average income (0.04), shorter hospital stays (-0.96), slightly lower additional charges (-0.69), lower rates of high blood pressure (0.29) and stroke (0.18), and average emergency admission rates (0.50). Cluster 2 consists of patients with above-average age (0.87), below-average income (-0.03), shorter hospital stays (-0.95), higher additional charges (0.81), higher rates of high blood pressure (0.54) and stroke (0.21), and higher emergency admission rates (0.51). Cluster 3 consists of patients with above-average age (0.87), slightly above-average income (0.008), longer hospital stays (0.95), higher additional charges (0.76), higher rates of high blood pressure (0.52) and stroke (0.20), and average emergency admission rates (0.50).

**F2. Discuss the results and implications of your clustering analysis**

I used the Elbow Method to find the optimal number of clusters for K-means by plotting the inertia against different values of k (number of clusters). The plot showed a noticeable "elbow" at k = 4, where the decrease in inertia slowed down significantly. This is why I chose 4 as the optimal number of clusters. The implication of this is that dividing the medical data into 4 clusters helps group patients into meaningful categories while keeping the analysis simple and efficient.

I created a scatter plot to visualize clusters using the variables Age and Initial\_days. The plot showed clear separation between clusters, with minimal overlap. The clusters were distributed across four quadrants. Cluster 0 represented younger patients with longer hospital stays. Cluster 1 represented younger patients with shorter hospital stays. Cluster 2 represented older patients with shorter hospital stays. Cluster 3 represented older patients with longer hospital stays. The implication of this is that the visualization confirms Age and Initial\_days are effective features for distinguishing patient groups. These clusters can help create targeted interventions based on age and hospital stay length.

The cluster summary table showed the mean values of continuous features and proportions of categorical features for each cluster. Cluster 0 consisted of younger patients with below-average income, longer hospital stays, lower additional charges, moderate rates of high blood pressure, and average emergency admission rates. Cluster 1 consisted of younger patients with higher income, shorter hospital stays, lower additional charges, lower rates of high blood pressure, and average emergency admission rates. Cluster 2 consisted of older patients with below-average income, shorter hospital stays, higher additional charges, higher rates of high blood pressure and stroke, and higher emergency admission rates. Cluster 3 consisted of older patients with above-average income, longer hospital stays, higher additional charges, higher rates of high blood pressure and stroke, and average emergency admission rates. The implication of this is that each cluster represents a distinct patient group with unique characteristics, helping the hospital understand its patients and plan resources and care effectively.

The hospital can use these results from the K-means clustering analysis to create targeted strategies that improve patient care and lower hospital costs. For cluster 0, which consists of younger patients with longer hospital stays, the hospital can focus on shortening stays through better discharge planning and follow-up care. They can also focus on preventive care, targeting issues like overweight and diabetes. For cluster 1, which consists of younger patients with shorter hospital stays, the hospital can focus on preventive care, though minimal intervention may be needed as this group likely includes low-risk patients. For cluster 2, which consists of older patients with shorter hospital stays, the hospital should prioritize high-risk older patients with conditions like high blood pressure, diabetes, and stroke, despite their shorter stays. For cluster 3, which consists of older patients with longer hospital stays, the hospital should implement chronic disease management programs to reduce complications and readmissions. High additional charges in clusters 2 and 3 offer opportunities to reduce costs through more efficient treatments and resource allocation. The hospital should prioritize these high-cost groups to reduce costs.

**F3. Discuss one limitation of your data analysis**

One limitation of our K-means clustering analysis is about the assumption of spherical clusters. K-means assumes that clusters are spherical and evenly distributed. However, real-world data like the medical dataset can form clusters that are non-spherical, unevenly sized, or overlapping. As a result, K-means may oversimplify the groupings and miss the data's true structure. Looking back at the cluster visualization from earlier, the clusters appear to be well-separated and compact. However, the visualization only uses two variables (Age and Initial\_days) and does not fully represent the multidimensional relationships in the medical dataset. Datasets that contain multiple features like the medical dataset often have irregular-shaped or overlapping clusters in the higher-dimensional space due to complex relationships between features.

**F4. Recommend a course of action for the real-world organizational situation from part B1 based on the results and implications discussed in part F2**

One recommended course of action for the hospital to address the question: "What distinct patient groups, based on demographic, health conditions and hospitalization details, can help the hospital create targeted strategies to reduce readmissions and lower hospital costs?" is to focus on cluster-specific interventions. For cluster 0, which consists of younger patients with longer hospital stays, the hospital should focus on improving discharge planning and follow-up care to reduce hospital stay duration. The hospital should also implement preventive care programs for issues like overweight and diabetes to reduce the likelihood of future complications and readmissions. For cluster 1, which consists of younger patients with shorter hospital stays, the hospital should emphasize preventive care, but with minimal intervention, as this group likely represents low-risk patients. The hospital should also explore ways to prevent unnecessary hospital visits for this group. For cluster 2, which consists of older patients with shorter hospital stays, the hospital should prioritize these high-risk older patients with conditions like high blood pressure, diabetes, and stroke, despite their shorter stays. The hospital should also allocate resources for early detection of chronic conditions to reduce readmissions. For cluster 3, which consists of older patients with longer hospital stays, the hospital should implement chronic disease management programs to reduce complications and readmissions. The hospital should also focus on optimizing care plans for long hospital stays to improve patient outcomes and lower costs.

**G. Panopto Presentation**

**Panopto video link:** <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=b687f937-0f8e-481d-ad8f-b275013fe2f8>

**H. Record the web sources used to acquire data or segments of third-party code to support the analysis. Ensure the web sources are reliable**

Yadav, A. (2024, July 14). *K-means clustering for dummies: A beginner’s guide*. Medium. <https://medium.com/@amit25173/k-means-clustering-for-dummies-a-beginners-guide-399fb8c427fd>

**I. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized**

Banoula, M. (2024, December 5). *All About K-means Clustering Algorithm*. Simplilearn.com. https://www.simplilearn.com/tutorials/machine-learning-tutorial/k-means-clustering-algorithm