International Journal of Chemical and Biochemical Sciences
(ISSN 2226-9614)

Journal Home page: www.iscientific.org/Journal.html

Exploring Optimal Cluster Quality in Health Care Data (HCD): Comparative Analysis utilizing k-means Elbow and Silhouette Analysis

T. Sumallika[1], V. Alekya[2], P.V.M. Raju[3], M.V.L.N. Raja Rao[4], D.E. Gnnana Shiney[5], M. Vijaya Sudha[6]

[1]*Assistant Professor, IT Department, Seshadri Rao Gudlavalleru Engineering College, India.*
[2]*Assistant Professor, IT Department, Anil Neerukonda Institute of Technology & Sciences, India.*
[3]*Assistant Professor, MBA Department, Seshadri Rao Gudlavalleru Engineering College, India.*
[4]*Professor, IT Department, Seshadri Rao Gudlavalleru Engineering College, India.*
[5]*Assistant Professor, IT Department, Seshadri Rao Gudlavalleru Engineering College, India.*
[6]*Assistant Professor, CSE Department, Sir C.R.R College of Engineering, India.*

Abstract

A dataset's natural grouping can be better understood with the help of clustering. Partitioning the data into logical groups will help them make sense of it. The methodologies employed and the finding of hidden patterns to ascertain the effectiveness of clustering. The goal of this research is to see if quality can be implemented in Clustering. The purpose of this research is to investigate the optimal features to use and compare the performance of clustering approaches. In this research, we compared the two feasible clustering techniques. Elbow and Silhouette Analysis, Here We considered two significant techniques to outline scores and disperse plots to recommend to detect anomalies, and afterward approve, the quality of clusters, Which is a proportion of the nature of a group, and utilized to track down the mean outline co-effective of the multitude of tests for a various number of clusters. The most elevated outline score shows the ideal number of groups. The experimental results over Elbow and Silhouette techniques using the outline score to decide the best worth of k for those informational indexes and optimal clustering. We also explored whether using any one of these two clustering techniques will give optimal clusters.

Keywords: Comparative cluster analysis, Health Care Data, Machine Learning, Elbow and silhouette techniques, Cluster Quality

Full length article **Corresponding Author*, e-mail: sumallika.p@gmail.com

1. Introduction

The strategy for recognizing comparable similarities of information in a dataset is called clustering. It is perhaps the most mainstream procedure in Machine Learning (ML). Data points in each group are relatively more like information points of that group than those of different groups [1]. In this article, we will be taking you through the upgraded grouping AI procedures and the examination of their quality. For the most part, we may classify the grouping into two different ways hard and soft clustering. In a hard grouping, each piece of information is either completely appropriate for a group or is not. Instead of assigning each data point to a distinct group, soft clustering assigns each data point a probability or chance of belonging to each cluster. The exploitation of Machine Learning can be seen in every domain, on Facebook, Twitter, and YouTube suggesting us a video depends on our set of experiences. Machine Learning is all over the place! [2]. Anomaly identification is the process of acknowledging and recognizing outlier [3] data in any data-based event or observation that varies substantially from the rest of the data. Financial scams, medical issues, e-commerce, or even malfunctioning equipment, healthcare suggestions, and judgments can all benefit from anomalous data. For Instance, shopping centers or shopping malls have regularly enjoyed the competition to expand their business by doing anomaly detection and henceforth making the right decisions to get tremendous benefits. To accomplish this assignment AI and ML are being useful by numerous supplies as of now. It is astonishing to understand the way that how AI can help with such desires. For instance, shopping malls make use of their customer's information and foster ML models to focus on the right ones. This expands deals as well as makes the edifices effective.

2. The working of the machine learning system

Supervised Learning and Unsupervised Learning are two types of machine learning. We train the computer in Supervised by giving it both independent and dependent variables, such as classifying or predicting values. Unsupervised learning is concerned with identifying the data's structures or patterns. Clustering, recommendation systems, and other algorithms in this category do not use labeled data (or the dependent variable is not there) [4]. Unsupervised Learning yields fantastic results, allowing anyone to infer a plethora of hidden relationships between various qualities or data. When it comes to detecting anomalies, the more data you have, the better results. The data must be labeled as good or poor for supervised machine learning. The features describe disease behavior, and patient behaviors are known as abnormal conditions. We divide features into five groups, each with hundreds or thousands of unique attributes: *a. Identity:* The number of digits in the patient's id, age factor, number of deviations in glucose levels, and smoke collectively gives the abnormal conditions. *b. Checkups:* The number and type of Checkups they made in the short period, Number and type of tests they have gone through will help in finding outlier behavior. A system of rules to follow while solving complex issues, such as a mathematical equation or a recipe, is known as an algorithm. Based on consumer information provided by our characteristics, like whether or not a transaction is fraudulent, the algorithm learns how to make predictions. We will begin by instructing the algorithm on the last several heart stroke data, Mall Customer data, and Bank Transactions data that we refer to as a training dataset. These training sets should contain as many anomalies as possible as the machine will benefit from them. After the training, we will have a model that is particular to the recommendation and can detect anomalies in milliseconds. We regularly monitored the model to ensure that it is operating correctly, and we are always looking for ways to improve it. Every record has a new model that we enhance, update, and upload regularly so that the system can always detect the most recent abnormality techniques.

3. Machine learning cluster methods

A cluster is a collection of objects belonging to the same class, or, in other words, objects with similar features are grouped in one cluster. It is the process of classifying objects into distinct groups, each of which contains objects that are quite similar to those in other groups. Simply grouping groups of people who have similar lots or activities and assigning them to clusters. In both labeled and unlabeled datasets, is used to connect patterns and structures [5]. Clustering is a type of exploratory data analysis technique that can discover leagues in data so that data points in the same league (cluster) are truly parallel to one another and data points in different clusters have different characteristics.

3.1. Definition of clustering

A clustering C means separating a data set into a groups of clusters Ci, i= 1,….H. A partitioning that maximizes distances among clusters and minimizes distance within clusters is commonly used definition of optimal clustering. There are various ways to designate the distances under and among clusters. Figure 2 demonstrates 5 important categories of clustering algorithms based on their characteristics.

3.1.1. Partitioning-based Clustering (PBC)

Clustering is based on the partitioning of variables into N numbers, each representing a cluster. This type of method targets benchmark similarity functions.
Example: K-means clustering, CLARANS(Clustering Large Applications depending upon Randomized Search)

3.1.2. Hierarchical-based Clustering (HBC)

Clustering can be done based on the hierarchy having a tree-type structure and categorized into two kinds *Agglomerative (bottom-up approach), and Divisive (top-down approach).* The agglomerative clustering method initially locates every position in a cluster and then amalgamates two points nearest to it. *(bottom-up approach)* in divisive clustering, the entire population is first taken into account as one cluster before being divided into smaller groups. *(Top-down approach).*Example: CURE (Clustering Using Representatives), BIRCH (Balanced Iterative Reducing Clustering and using Hierarchies).

3.1.3. Density-based Clustering (DBC)

This type of clustering method focuses on the dense region that has some similarities; these will provide ample precision and also the capability to amalgamate two clusters. Most Probable forming cluster shape in this method is spherical hence it is hard to see arbitrarily shaped clusters. Example: DBSCAN (Density-based Spatial Clustering of Applications with Noise), OPTICS (Ordering Points to Identify Clustering Structure).

3.1.4. Grid-based Clustering (GBC)

The data is divided into a limited number of cells using a grid-like configuration in this method. On such grids (i.e., quantized space), several number-quick clustering techniques that don't depend on the quantity of data objects are applied [6]. Example: STING (Statistical Information Grid), Wave cluster, CLIQUE (Clustering In Quest).

3.1.5. Model-based Clustering (MBC)

These methods utilize a prevailing mathematical model to fit and standardize the data, presuming the data is in the form of possibility distributions, and use standard statistics to figure out how many clusters there are. To account for noise and outliers, the standard statistics for robust clustering are computed. In terms of how they create clusters, these clustering techniques can be split into two groups: statistical and neural network approaches. Probability metrics are used for cluster selection in model-based statistical algorithms, while unit-carrying weights for input and output are used in

neural network algorithms. Example: MCLUST (Model-based Clustering), GMM (Gaussian Mixture Models).

3.2. Applications of Clustering in Machine Learning

Numerous real-world issues, such as cancer cell identification, social network analysis, market and customer fragmentation, search result clustering, recommendation systems, and social network assessment all use clustering techniques.[7]. They concentrated on data types, technique biases to a priori parameters, cluster form, outlier presence, the volume of data, dimension of data, Missing Values, etc. to find the most efficient and efficient approach for optimal clustering. Traditional clustering algorithms primarily fall into two categories: partitioned and hierarchical [8].

4. Related Work

In this paper, we aim to discuss the comparison of the performance of cluster quality between the Elbow Technique and Silhouette analysis. The K-means clustering technique is an example of operating anomaly identification utilizing machine learning. Depending on their plotted distance from the nearest cluster, this technique is used to identify the outlier. K-means clustering [9] forms numerous clusters of averaged data points. The nearest mean value belongs to the cluster of objects. An outlier is any object with a threshold higher than the mean of the most recent clusters. The Davies-Bouldin (DB) index [10-11] assesses the distribution of data formulated on the distances among cluster centroids, taking into account both the greatest intra-cluster range and the minimum inter-distance.

4.1. Elbow Technique

Unsupervised K-means clustering is a simple and popular ML algorithm. There are two ways that we can assess the algorithm. The silhouette technique and the elbow technique are the two. The elbow is a very straightforward mechanism that gives us a plot with an elbow-like shape [12]. The graph allows us to infer the ideal value of k with ease. As the plot can be hazy at times, perhaps we will become unclear when we encounter an intricate one. Moreover, by using the silhouette approach, we can quickly determine the precise value of k and determine the silhouette coefficient [13].

4.2. Silhouette Analysis

The degree of detachment among clusters can be found using silhouette assessment [14]. For every sample: Calculate the mean separation between all the data points in the same cluster (ai) [15-16]. Calculate the mean separation between all the data points in the closest cluster (b^i). And the coefficient [17]. The coefficient can take values in the interlude [-1, 1] [18]. A value of +1 for the Silhouette coefficient indicates that the sample is remote from the nearby clusters. [5]. When the Silhouette coefficient is 0, the sample is on or very close to the decision boundary between two neighboring clusters [19]. A silhouette coefficient of 0 indicates that a sample is an outlier or that it may have been misallocated to a cluster [20**].** Therefore, we hope that the coefficient is as large as possible and close to 1 to obtain good clustering. We used mall customer's data set here to run a profile analysis and it is obvious that sets of data points.

$$\frac{b^i - a^i}{max(a^i, b^i)}$$

5. Methodology

Here the Methodology follows six steps, in the Jupiter platform; different datasets are downloaded from Kaggle. The following figure shows the steps involved in the methodology of Comparative Analysis of Cluster Quality on Health Care Data.

Steps Involved in the Methodology

Step 1: Import all the necessary Python modules into our ML program
Step 2: To save the data to a pandas data frame, read the data from an Excel file. Source Data.
*Step 3: T*rain the machine learning algorithm utilizing the past data point gathered from HCD (Health Care Data).
*Step 4: A*pply algorithm K-means cluster to preprocess the data
Step 5: Apply Elbow Method and Silhouette Analysis on HCD (Health Care Data).
Step 6: Analyze the results

Step 1: Import the required modules

Our machine learning system was first proffered with all the requisite Python modules. These libraries and modules won't be detailed. import pandas as pd from sklearn. Preprocessing import StandardScaler. Pandas is a Python library used for reading data from various sources for assessment and manipulation. Pandas make it simple to work with time-series data and access and create data in file formats like CSV, Excel, and others. Python also has a library called Scikit-learn. It offers straightforward and effective tools for the analysis of predictive data. For the same, we will import particular Scikit Learn modules.

Step 2: Read the data source

The data points will then be read from an Excel file and saved to Source Data in a pandas data frame.
SourceData=pd.read_excel("Stroke.xlsx") # Fill Pandas DataFrame with training data.Testdata = "pd.read_excel("Predict.xlsx")" # Include test data.
Note: I've saved the Python code, the Excel sheet with the historical data points, and the future data set all in the same folder. As a result, just the file names are supplied in the code. If files are spread across many folder locations, complete paths must be given.

Table 1. Attributes used in three different datasets

Stroke Dataset Attributes (Diabetes)	Stroke Dataset Attributes (Alcohol)	Stroke dataset Attributes (smoking)
Number of pregnancies	Age	Id
Glucose level	Sex	gender
Blood pressure	Hypertension	age
Ski thickness	Alcohol overuse	hypertension
Insulin	Heart failure	heart_disease
BMI	Atrial fibrillation	ever_married
Age	COPD	work_type
Job type	Hyperlipidemia	Residence_type
Married	Job type	avg_glucose_level
Gender	Ever married	BMI
Stroke	Stroke	smoking_status
Outcome	Coronary artery diseases	stroke

Table 2. Comparative Analysis of cluster quality

	Health care Data Set (no.of.Clusters)		Stroke diabetes data Set(no.of.Clusters)		Stroke Alcohol Consumption Data Set (no.of.Clusters)	
	Elbow	**Silhouette**	**Elbow**	**Silhouette**	**Elbow**	**Silhouette**
Same	✓		✓		✓	
Different	--		--		--	

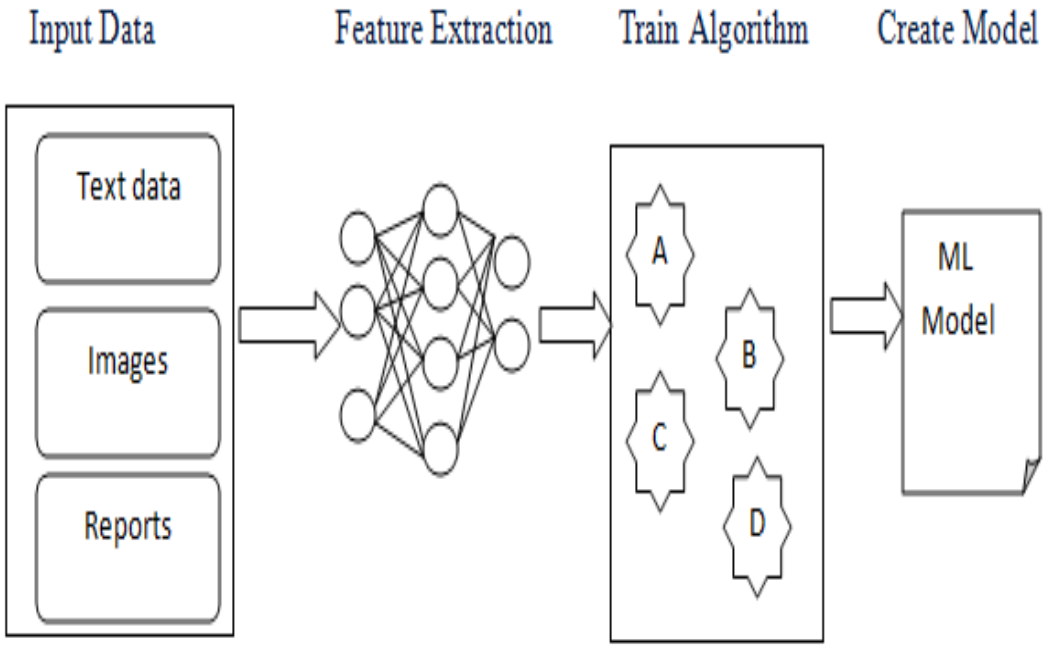

Figure 1. Structure of ML System

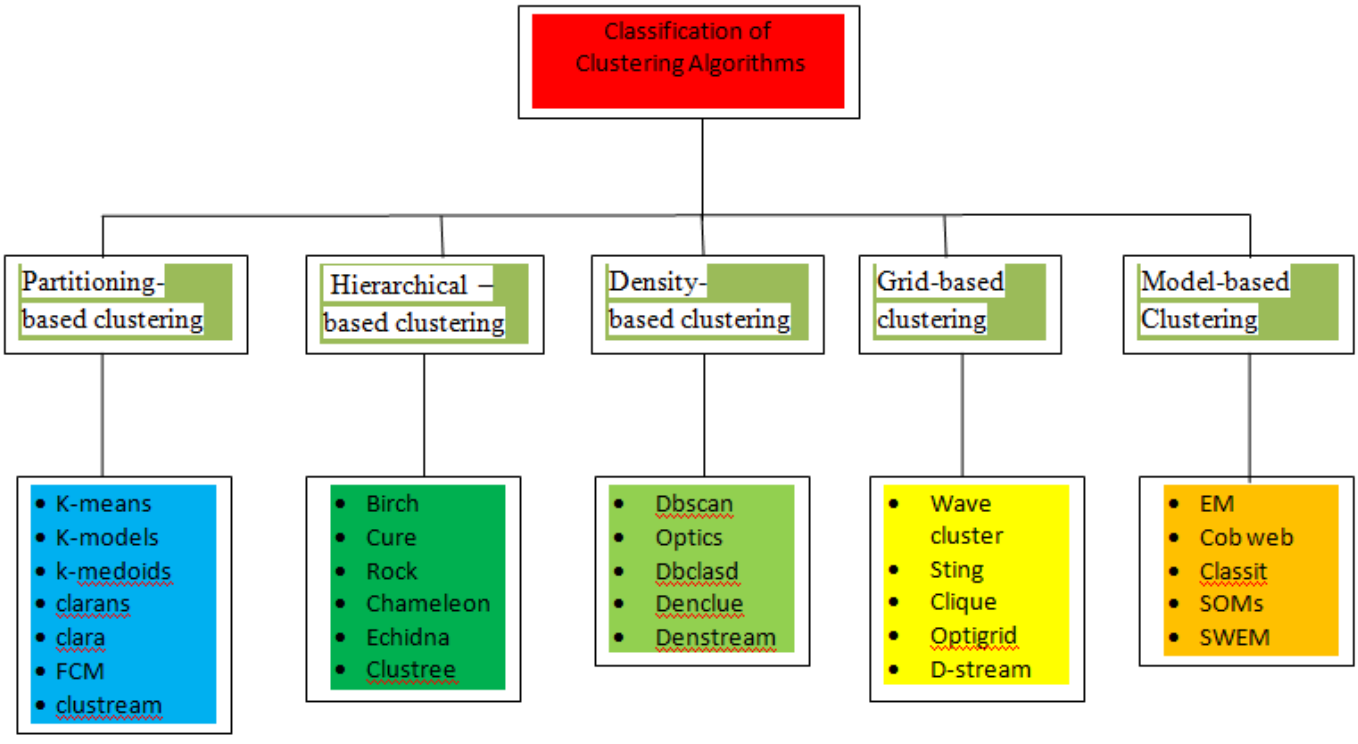

Figure 2. Various types of clustering methods

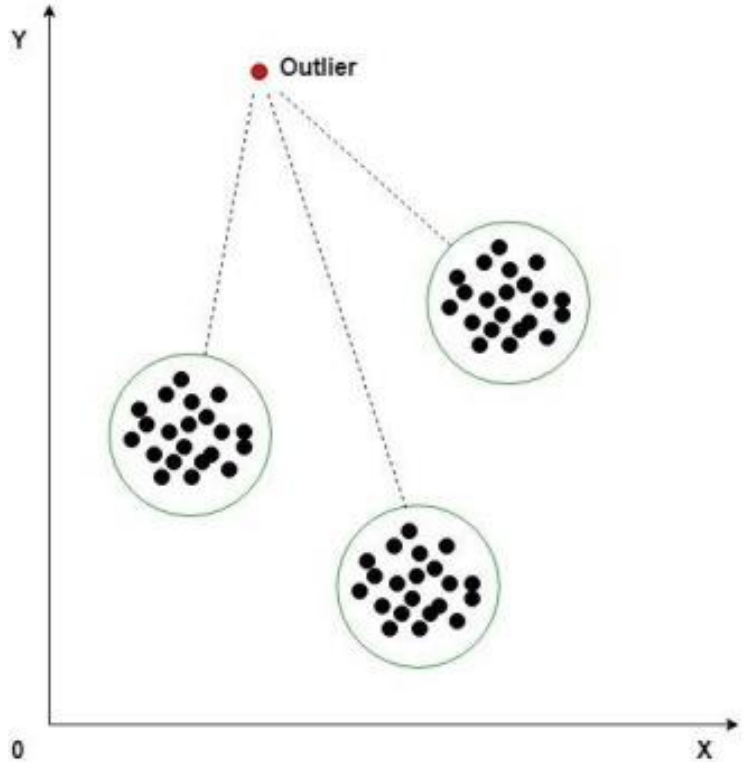

Figure 3. Multiple Clusters in K-means

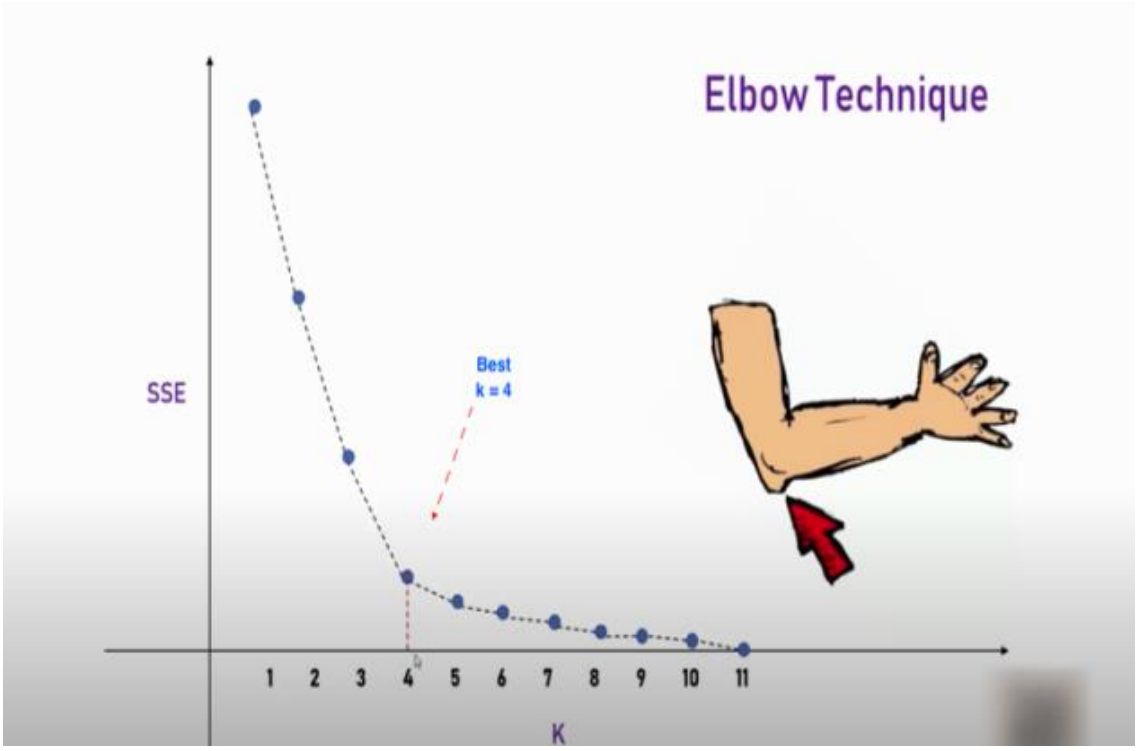

Figure 4. Optimal number of clusters using the Elbow method

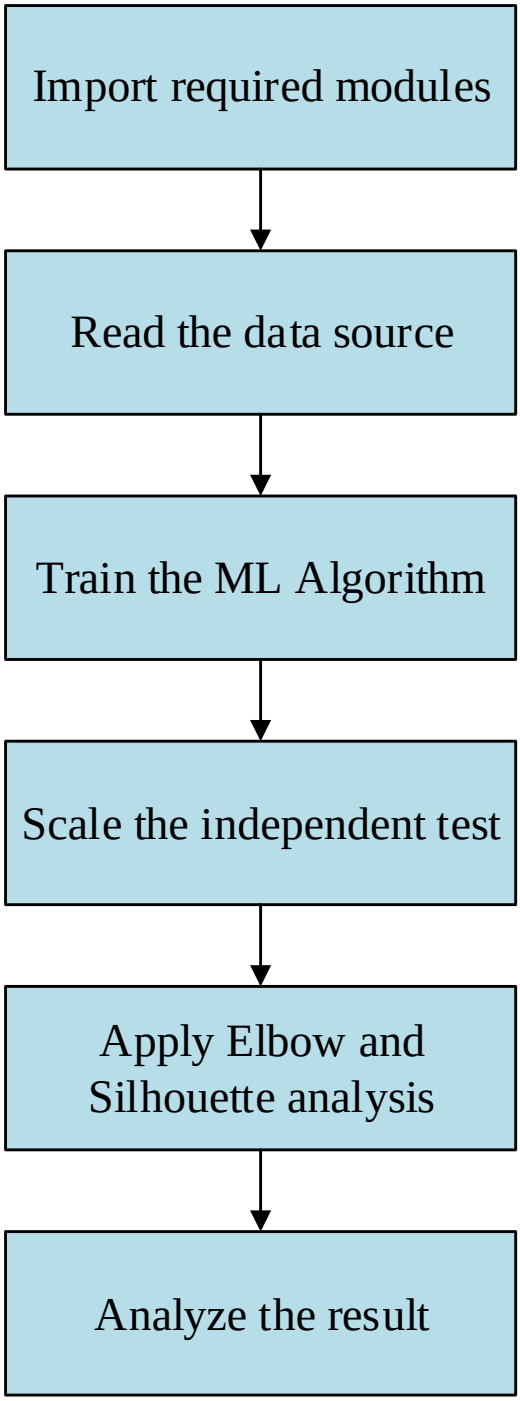

Figure 5. Flowchart of the methodology

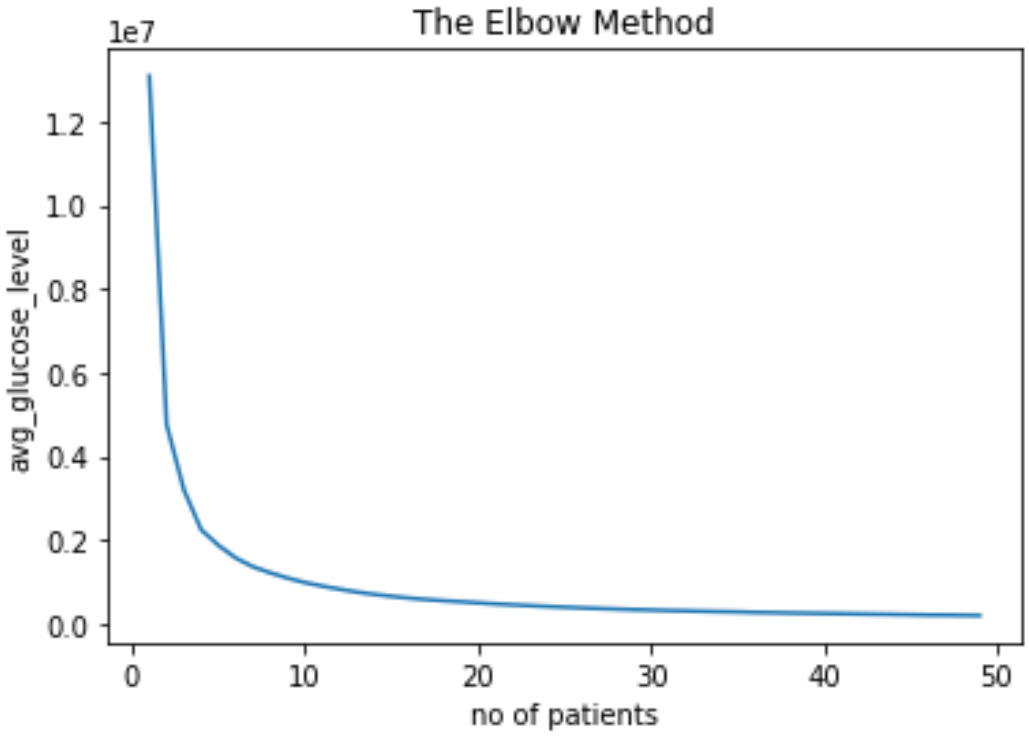

Figure 6. Number of Clusters utilizing Elbow Methodology

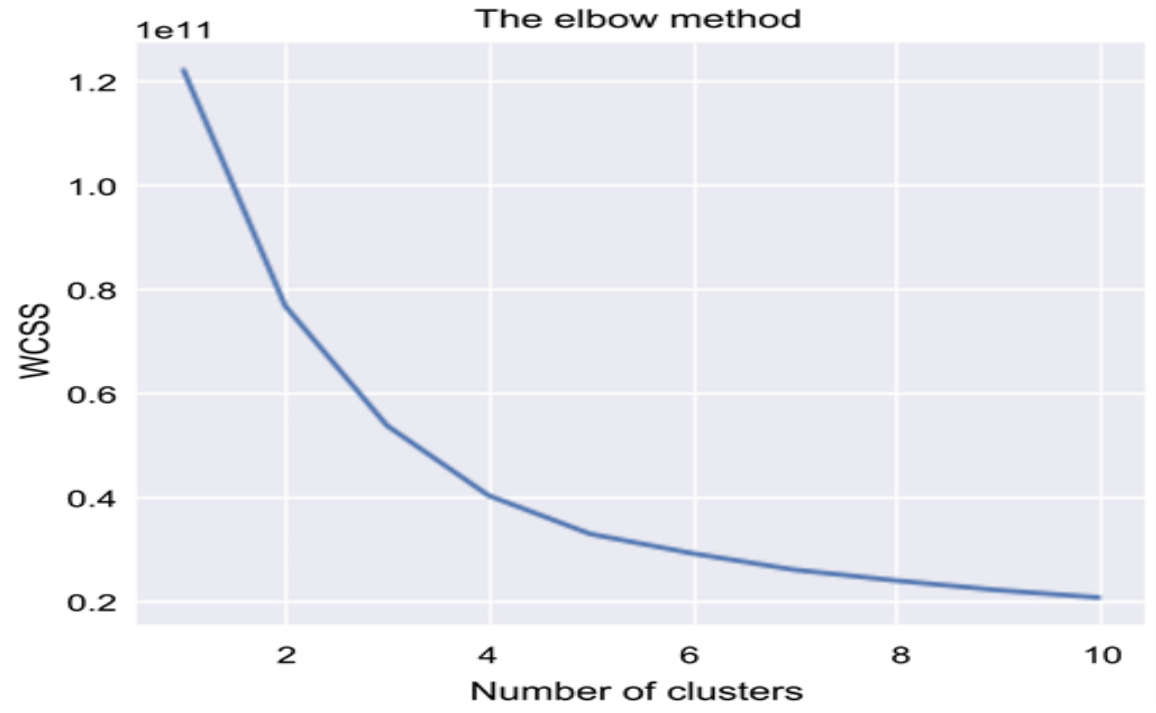

Figure 7. Elbow method

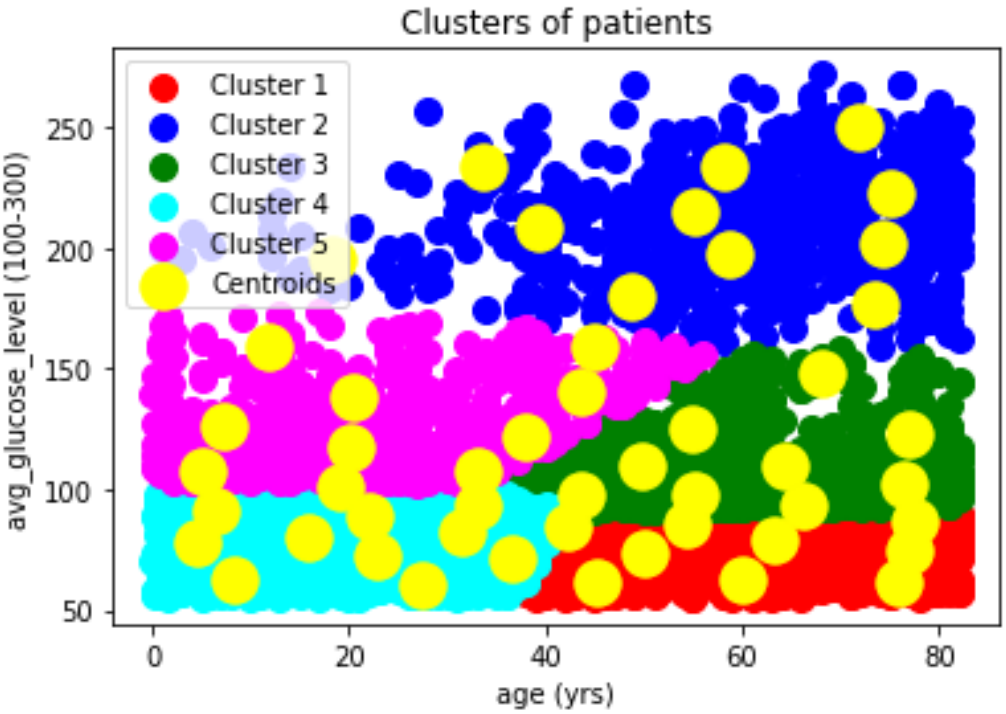

Figure 8. Clusters plot using Elbow Method

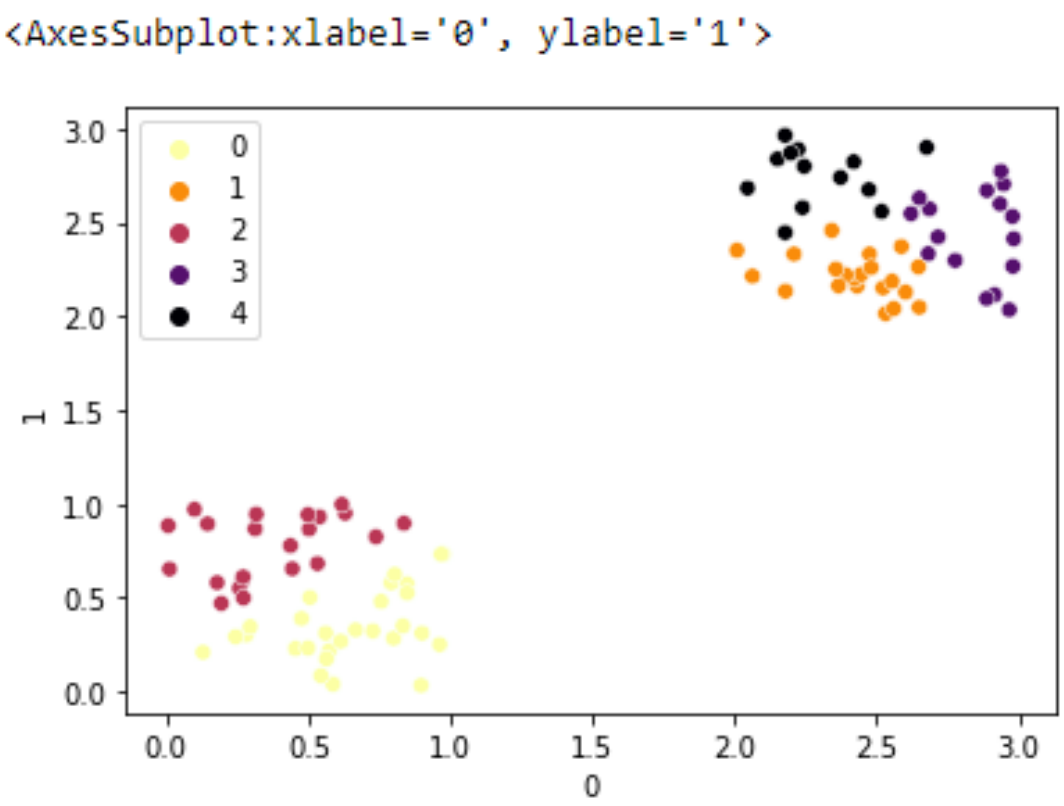

Figure 9. Number of Clusters using the Silhouette technique

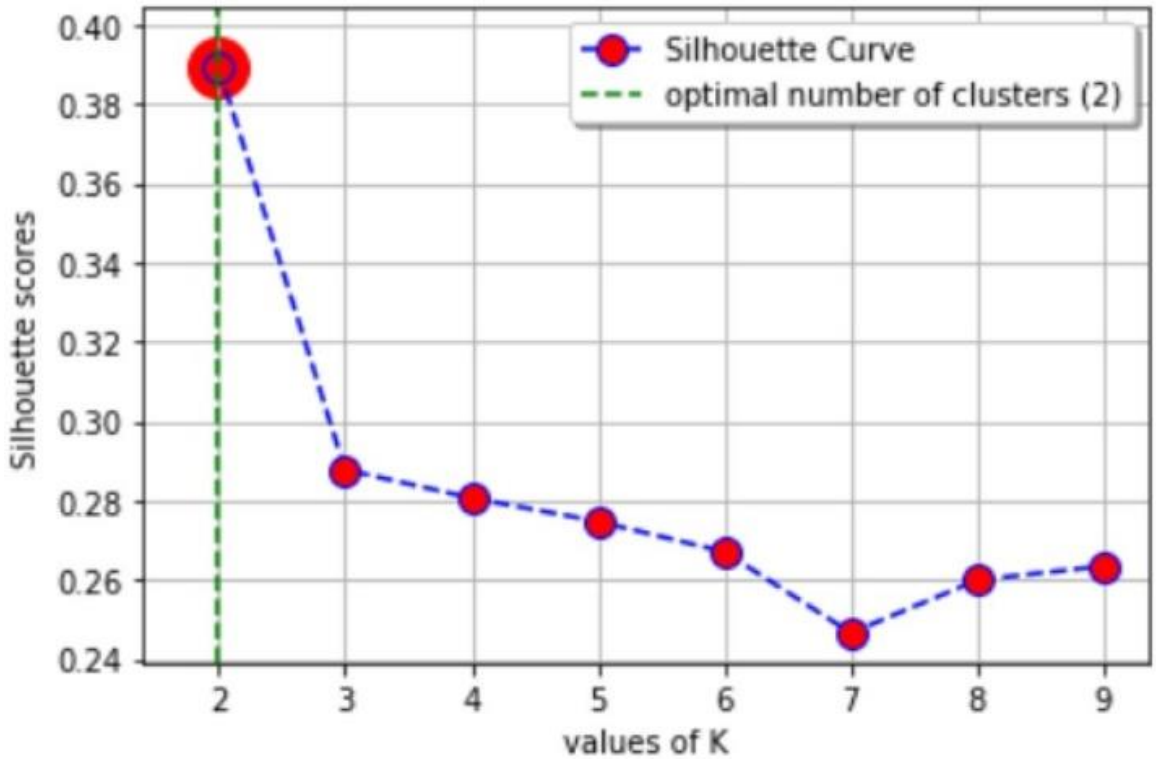

Figure 10. Plotting the Silhouette scores for different K values

Step 3: train the machine learning algorithm using the past data point collected from HCD (Health Care Data)

Now train the data points from the data set. Considerable four data points are Gender, age, hypertension, heart disease, etc from Excel into Pandas Data frame. We will utilize Gender, age, hypertension, heart disease, etc to predict the occurrence of stroke in the future. The projected variable is the dependent variable, whereas the independent variables are the inputs used to make the prognostication. To train the ML algorithm, we will be utilizing the previously gathered data points,

The variables SourceData_train_independent and SourceData_train_train_dependent.

SourceData_train_independent= SourceData.drop(["Positive prediction"], stroke=1) datasetSourceData_train_dependent=SourceData["Positive prediction"].copy()

Step 4: Scale the data from the independent test and training

The age value ranges from months to 82 years, whereas the glucose level varies from 55.12 mg/dL. to 271.74 mg/dLsmoking status values vary like never smoke, unknown, formerly smoked, smokes. Because the ranges of the independent variables are so wide, scaling is necessary to prevent an unexpected influence of one variable in this case, the value of age, glucose levels, or smoking status over other variables. Only the independent variables need to be scaled. The independent train and test variables are scaled and given the names X-train and X_test, respectively, in the code below. We don't scale the dependent trained variable when we save it in y_train.

```
X_train=
sc_X.fit_transform(SourceData_train_independent.values);
sc_X = StandardScaler()
X_test=sc_X.transform(Testdata.values)
y_train=SourceData_train_dependent
```

Step 5: apply Elbow Method and Silhouette Analysis

The ML model will then be trained using X_train and y_train, respectively, are independent and dependent train data points. To clear up any misunderstandings, we first used a fit methodology to fit the X_train and y_train in the code below. Then, using the X_test variable, we pass the test independent variable values for “age”, “avg glucose level” and “smoking status” and then put the prediction in the "stroke" variable.

```
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
age = []
for i in range(1, 16):
   kmeans = KMeans(n_clusters=i, init='k-means++', random_state=0)
```

```
    kmeans.fit(X)
    age.append(kmeans.inertia_)
plt.plot(range(1, 16), age)
plt.title('The Elbow Method')
plt.xlabel('Number of Patients')
plt.ylabel('Age')
plt.show()
```

Step 6: Analyze the results

Analysis can be done by plotting clusters with customized colors, it will provide transparent and clear results for analyzers to analyze the data for future recommendations see Fig 6.

```
import matplotlib.pyplot as plt
import seaborn as sns
plt.plot(range(1, 16), age)
plt.title('The Elbow Process')
plt.xlabel('Number of Patients')
labels = KMean.predict(z)
silhouette_score = silhouette_score(z, labels)
print(f"Silhouette Score (n=4): {silhouette_score}")
sns.scatterplot(z[0], z[1], hue=labels, palette='inferno_r')
```

6. Implementation

6.1. Clustering using the Elbow Method

The diminishing of a sum of squared distances has a slower rate after K=2, as illustrated in Figure 6, leading one to the conclusion that K=2 is the ideal value. These two clustering techniques are then used to group the patients into two groups—high priority and low priority. The Silhouette coefficient is used to compare the accuracy of these two clustering techniques. The silhouette coefficient compares how closely a point resembles its cluster in relation to other clusters.

```
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
kmeans_model = KMeans(n_clusters=5, init='k-means++', random_state=0)
y_kmeans = kmeans_model.fit_predict(X)
plt.scatter(X[y_kmeans == 0, 0], X[y_kmeans == 0, 1], s=110, c='red', label='Cluster 1')
plt.scatter(X[y_kmeans == 1, 0], X[y_kmeans == 1, 1], s=110, c='blue', label='Cluster 2')
plt.scatter(X[y_kmeans == 2, 0], X[y_kmeans == 2, 1], s=110, c='green', label='Cluster 3')
plt.scatter(X[y_kmeans == 3, 0], X[y_kmeans == 3, 1], s=110, c='cyan', label='Cluster 4')
plt.scatter(X[y_kmeans == 4, 0], X[y_kmeans == 4, 1], s=110, c='pink', label='Cluster 5')
plt.scatter(kmeans_model.cluster_centers_[:, 0], kmeans_model.cluster_centers_[:, 1], s=320, c='yellow', label='Centroids')
plt.title('Clusters of Patients')
plt.xlabel('Age (years)')
plt.ylabel('Average Glucose Level (100-300)')
plt.legend()
plt.show()
```

From 2 to 4 was where the elbow curvature began (figure 7). this implies that a dataset may contain two to four clusters. We chose K = 2 as the first K value because we anticipate seeing two big groupings. The programme then determines how far (or similar) each data point is to each of the other K points. Every data point is assigned to one of the K clusters based on these distance values. In other words, the cluster whose centroid is closest to the data point among all other similar K centroids is the cluster to which the data point is assigned. The centroids of each of the K clusters are recalculated once all of the data points have been assigned to one of the K clusters. The procedure is then repeated using the new centroids as the cluster centre. Up until there are no more cluster assignments made, this iterative process is repeated. Each data point's squared distance, cluster centroid, and residual sum of squares (RSS) were added together as a criterion for termination.

Analyzing the results (according to the Elbow method)

From the fig 8 cluster, 5(pink colored) we can observe that people with middle age and with moderate glucose levels are quite reasonable. In cluster 4(cyan colored) we can notice that people in the early stages having low glucose levels may be recommended for doctor consultation. In cluster 3(green colored) we can observe that people above middle age and with moderate glucose levels are quite good enough. In cluster 2 (blue-colored) we observe that people above middle age with high glucose levels are in the danger zone and are highly recommended for treatment. In cluster 1 (red-colored) we observe that the people above middle age with very low glucose levels are also in the danger zone and are highly recommended for treatment. Maybe these are the different groups of people who required their health care recommendations. So, healthcare professionals can take decisions for the sake of patients. As a final point, based on our ML clustering technique we may assume that to increase the betterment of healthcare services, healthcare professionals should target people belonging to Cluster 4, cluster 2, and cluster 1 to increase their life span. In conclusion, Table 2 contrasts the quality of the clusters (CQ) with Elbow Technique and Silhouette Score.

6.2. Clustering using silhouette score

```
import numpy as np
import pandas as pd
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
import seaborn as sns
a = np.random.rand(50, 2)
y = 2 + np.random.rand(50, 2)
z = np.concatenate((a, y))
z = pd.DataFrame(z)
kmeans = KMeans(n_clusters=5)
kmeans.fit(z)
labels = kmeans.predict(z)
silhouette_score_value = silhouette_score(z, labels)
print(f"Silhouette Score (n=5): {silhouette_score_value}")
sns.scatterplot(z[0], z[1], hue=labels, palette='inferno_r')
```

Results

```
Silhouettee Score(n=4):0.5114553522196848
Silhouettee Score(n=5):0.498396561386838
Silhouettee Score(n=6):0.51110052456428703
```

Silhouettee Score(n=7):0.5543224440106885
Silhouettee Score(n=8):0.5485495207695185
Silhouettee Score(n=9):0.53095189978940807
Silhouettee Score(n=10):0.49744974722721493
Silhouettee Score(n=11):0.5787058694887515
The number of cluster values using the Silhouette technique is shown in Figs.9 for K = 2 and K = 3, respectively. From Fig. 9 it is shown that anomalies (red square) are far from the clusters and have a small effect in detecting anomalies. But in Fig.9, it is shown that anomalies are mixed with the normal data instances, and this phenomenon may hamper detecting anomalies. Figure 10 illustrates the silhouette scores for different K values. From Fig. 10, it is observed that cluster value K of 2 has the highest Silhouette score.

Analyzing the results (according to the Silhouette technique)

The optimal number of clusters can be calculated in this methodology, the experimental results show that the n value(number of clusters) can be observed by the silhouette score. We got the numerical value which tells the number of clusters to be formed for n=4,5,6…. the silhouette score is 0.5114…, 0.4983…, 0.5111…etc respectively, which is approximately equal to 5 i.eThe optimal number of clusters is 5. See Table 1 for a Comparative Analysis of cluster quality.

7. Datasets used and tested

Three datasets Stroke, diabetes, and Alcohol consumption datasets are tested with various numbers of attributes to find the quality of clustering (CQ). Elbow and Silhouette clustering techniques are applied to these datasets and analyzed their scores. Table 1 gives information on attributes in each dataset. The above techniques involve using various libraries and functions to perform clustering analysis and evaluate the clustering results. Let's analyze each technique used in the provided code:

7.1. Importing Libraries

Import numpy as np: Imports the NumPy library, which provides support for numerical operations and array manipulation.
Import pandas as pd: Imports the pandas library, which provides data structures and data analysis tools.

7.2. Clustering Technique

From sklearn.cluster import KMeans: Imports the KMeans class from the sklearn.cluster module. K-means clustering is a popular unsupervised learning algorithm that partitions data into a specified number of clusters based on similarity.

7.3. Generating Data

a = np.random.rand(50, 2): Generates a random 50x2 array of values between 0 and 1 using NumPy.
y = 2 + np.random.rand(50, 2): Generates a random 50x2 array of values between 2 and 3 using NumPy.
z = np.concatenate((a, y)): Concatenates the arrays a and y along the rows to create a combined dataset.

7.4. Data Manipulation

z = pd.DataFrame(z): Converts the NumPy array z into a pandas DataFrame for easier data handling and analysis.

7.5. Clustering Analysis

kmeans = KMeans(n_clusters=5): Creates a KMeans object with 5 clusters.
kmeans.fit(z): Performs the K-means clustering on the dataset z using the fit() method.
labels = kmeans.predict(z): Assigns cluster labels to each data point in z using the predict() method.

7.6. Evaluation

silhouette_score_value = silhouette_score(z, labels): Computes the silhouette score, a measure of how well each data point fits into its assigned cluster.
print(f"Silhouette Score (n=5): {silhouette_score_value}"): Prints the silhouette score for the clustering result.

7.7. Visualization

sns.scatterplot(z[0], z[1], hue=labels, palette='inferno_r'): Creates a scatter plot of the data points in z with different colors representing different clusters. These techniques collectively allow for data generation, clustering analysis, evaluation, and visualization of the clustering results. The K-means algorithm is used to group data points into clusters, and the silhouette score provides an evaluation of the clustering quality. The scatter plot visualization helps visualize the cluster assignments. Overall, these techniques provide a pipeline for performing K-means clustering and analyzing the results using the provided dataset.

8. Conclusion

In this research, we presented a Comparative Analysis of Cluster Quality on Health Care Data (HCD) using k-means Elbow and Silhouette Analysis. There are several studies and opinions for the optimal clustering methods in Machine Learning (ML), Data Analysis through clustering to detect outliers ML can help us detect, classify, and segment diseases from some test results. It also provides a quick comparison between clustering techniques and Machine Learning techniques with priorities of figuring out the most efficient and productive technique for optimal clustering (cluster quality).In this research, we have implemented both Elbow Method and Silhouette Analysis on HCD (Health Care Data)for various data sets. The experimental results say that when both the Elbow method and Silhouette analysis are applied for any type of data set the optimal clusters, give the same number of clusters. e. cluster quality (CQ). Here we presented one prominent experimental result which gives the same number of clusters hence one can be able to use any one of these two methods. Our methodology produces real-time datasets (which are not excellently dispersed) from which the ideal number of clusters can be determined through analysis. To sum up, in this experiment, I instruct the model repeatedly while running the code to determine the outcome for fresh cases. Although this research provides promising results in Cluster Quality for large datasets and complicated examples, this might not be feasible.

References

[1] J. MacQueen. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. 1 (14) 281-297.

[2] S.M. Sasubilli, A. Kumar, V. Dutt. (2020). Machine learning implementation on medical domain to identify disease insights using TMS. In 2020 International Conference on Advances in Computing and Communication Engineering (ICACCE). 1-4.

[3] G. Singh, R. Gupta, A. Rastogi, M.D. Chandel, R. Ahmad. (2012). A machine learning approach for detection of fraud based on svm. International Journal of Scientific Engineering and Technology. 1 (3) 192-196.

[4] Y. Kumar, K. Kaur, G. Singh. (2020). Machine learning aspects and its applications towards different research areas. In 2020 International conference on computation, automation and knowledge management (ICCAKM). 150-156.

[5] H. Kim. (2020). Performance analysis of K means clustering algorithms for mMTC systems. In 2020 International Conference on Information and Communication Technology Convergence (ICTC). 30-35.

[6] C.F. Tsai, H.C. Wu, C.W. Tsai. (2002). A new data clustering approach for data mining in large databases. In Proceedings International Symposium on Parallel Architectures, Algorithms and Networks. I-SPAN'02. 315-320.

[7] F. Wang, H.H. Franco-Penya, J.D. Kelleher, J. Pugh, R. Ross. 2017). An analysis of the application of simplified silhouette to the evaluation of k-means clustering validity. In Machine Learning and Data Mining in Pattern Recognition: 13th International Conference, MLDM 2017, New York, NY, USA, July 15-20, 2017, Proceeding. 13 291-305.

[8] C. Bishop. (2006). Pattern recognition and machine learning. Springer google schola. 2 5-43.

[9] D.L. Davies, D.W. Bouldin. (1979). A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence. (2) 224-227.

[10] A. Et-taleby, M. Boussetta, M. Benslimane. (2020). Faults detection for photovoltaic field based on k-means, elbow, and average silhouette techniques through the segmentation of a thermal image. International Journal of Photoenergy, 2020. 1-7.

[11] I. Polian, S.M. Reddy, B. Becker. (2008). Scalable calculation of logical masking effects for selective hardening against soft errors. In 2008 IEEE Computer Society annual symposium on VLSI. 257-262.

[12] M. Maniatakos, Y. Makris. (2010). Workload-driven selective hardening of control state elements in modern microprocessors. In 2010 28th VLSI test symposium (VTS). 159-164.

[13] Y. Yu, B. Bastien, B.W. Johnson. (2005). A state of research review on fault injection techniques and a case study. In Annual Reliability and Maintainability Symposium, 2005. Proceedings. 386-392.

[14] A. Evans, M. Nicolaidis, S.J. Wen, T. Asis. (2013). Clustering techniques and statistical fault injection for selective mitigation of SEUs in flip-flops. In International Symposium on Quality Electronic Design (ISQED). 727-732.

[15] A.K. Jain, R.C. Dubes. (1988). Algorithms for clustering data. Prentice-Hall, Inc..

[16] N.S. Altman. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician. 46 (3) 175-185.

[17] J. Fan, Q. Zhang, J. Zhu, M. Zhang, Z. Yang, H. Cao. (2020). Robust deep auto-encoding Gaussian process regression for unsupervised anomaly detection. Neurocomputing. 376 180-190.

[18] I.H. Sarker, A.S.M. Kayes, P. Watters. (2019). Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. Journal of Big Data. 6 (1) 1-28.

[19] P. Bholowalia, A. Kumar. (2014). EBK-means: A clustering technique based on elbow method and k-means in WSN. International Journal of Computer Applications. 105 (9).

[20] F. Liu, Y. Deng. (2020). Determine the number of unknown targets in open world based on elbow method. IEEE Transactions on Fuzzy Systems. 29 (5) 986-995.