**International Journal of Management, IT & Engineering**
Vol. 13 Issue 08, August 2023
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

# Enhancing Healthcare Insights, Exploring Diverse Use-Caseswith K-means Clustering

**Sameer Shukla***

**Abstract**

K-means clustering algorithm is a unsupervised machine learning algorithm that partitions data into k clusters, it's an iterative algorithm where in each observation belongs to the cluster with the nearest mean (centroids). The K-means clustering can be extremely useful in healthcare sector, the paper explores two compelling use-cases to demonstrate how impactful it can be for healthcare professionals. The first use-case involves clustering patient data based on specific features, enabling the identification of distinct patient subgroups with shared symptomps. The second use-case explores medical image segmentation, medical images can be partitioned efficiently and accurately using K-means. The results highlight the significance of K-means in enhancing the healthcare insights and facilitates data-driven decision making. The research showcases the remarkable potential of K-means clustering in healthcare domain.



***Keywords:***

K-means clustering.
Image segmentation.
Healthcare insights.
Elbow method.
AutoML.
Machine Learning.
Google Cloud Platform.

***Author correspondence:***

Sameer Shukla,
Lead Software Engineer, Irving, TX, USA
Email: sameer.shukla@gmail.com

## 1. Introduction

Clustering is a technique used in data mining to group similar objects into clusters, its an unsupervised machine learning algorithm that means it can classify data without having to been trained first with labeled data.

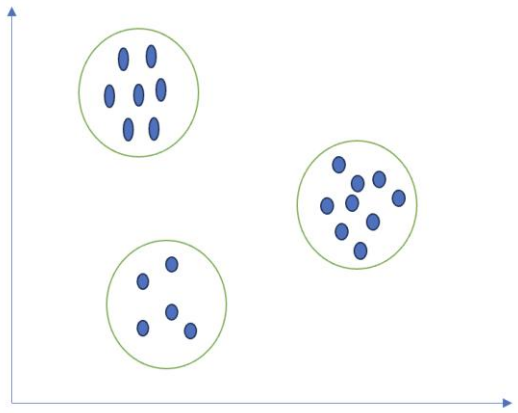

Figure 1. *Clustering*

K-means is straightforward algorithm with simple steps that includes, K-means [1] is a clustering algorithm that groups data points into k clusters. If the dataset contains features with different scales (e.g one feature has values

* Doctorate Program, Linguistics Program Studies, Udayana University Denpasar, Bali-Indonesia (9 pt)

**International Journal of Management, IT & Engineering**
Vol. 13 Issue 08, August 2023
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

in the range of 0 and 1 and other feature has values in the range of 1000 to 10000) then this data needs to be scaled first, this ensures that all the features contribute to the clustering process and the algorithm is not dominated by one specific feature. If the dataset contains categorical variables these variables need to be one-hot [2] encoded, encoding converts categorical values to binary vectors. After pre-processing K-means algorithm begins by selecting the number of clusters (k) and randomly initialize the cluster centres, for each data point in the dataset, algorithms required to calculate the distance between each cluster center and assign the data point to the cluster whose center is closest to it.

After assigning all data points to the clusters, algorithm requires to calculate the new cluster centres, then calculate the mean of all data points assigned to each cluster this calculated mean will be the new centroid of the cluster. The biggest challenge in the K-means algorithm is to find the optimal number of clusters (k) in the dataset, to identify the optimal number of clusters an elbow method [3] should be utilized, point to remember is adding more clusters does not reduce the intertia.

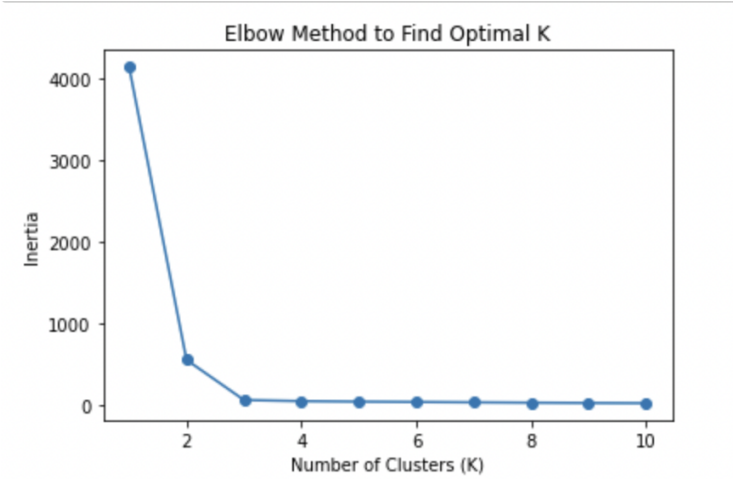


Figure 2. *Elbow Method*

K-means is a simple and intuitive algorithm but can be very effective for data clustering, it's simple and easy to-understand algorithm and it's computationally efficient and effective for huge datasets. K-means is most effective when dealing with spherical or isotropic clusters, as it may struggle with non-convex or irregularly shaped clusters. For such cases, other clustering algorithms like DBSCAN [4] or Gaussian Mixture Models might be more appropriate.

**2. LiteratureReivew**

K-means clustering is simple yet powerful algorithm for data analysis in healthcare domain, facilitating insights and improvements in various aspects of medical research and patient care.

As per "Haratyetal. (2015)" [5] using K-meanstoidentify patterns in healthcare data can help in improving disease diagnosis and treatment, drug discovery and healthcare policy making.

Several other research have demonstrated the utility of K-means in efficient healthcare resource management, studies have used clustering techniques to group hospitals or medical facilities based on patient characteristics, disease commonness. This information assists policy makers in strategic planning, improving resource distribution and enhancing the overall healthcare system efficiency.

The researchGabrieleSpini, etal. [6] proposes a secureclustering algorithm for hospital workflow optimization. The research proposes a new clustering algorithm, secure K-means that is designed for hospital workflow optimization. Secure k-Means can be used to cluster hospital staff based on their activity patterns. This can be used to optimize the workflow of the hospital and to improve patient care. Secure k-Means can be used to cluster hospital transactions based on their characteristics. This can be used to detect fraudulent transactions and to protect the hospital from financial losses.

In the field of medical imaging, K-means clustering has proven effective for image segmentation tasks. By partitioning medical images into regions of interest, such as tumors, tissues, radiologists can obtain more accurate and detailed information for diagnosis and treatment planning. K-means-based image segmentation enables faster andmoreautomatedanalysisofmedicalimages, reducingtheburdenonhealthcareprofessionals.

The literatureon K-meansclustering in thehealthcaresectorshowcasesitsvariety and importantcontributionsin healthcare sector. From patient data analysis and disease subtyping to medical image segmentation and healthcare resource management, K-means clustering continues to play a crucial role in advancing healthcare insights, patient care, and overall healthcare system efficiency. Cloud providers came up with a new advancement of AutoML where one need to be an expert in machine learning can utilize the powerful cloud tool for the contribution of medical advancement.

## 3. ResearchMethod

K-means clustering algorithm can be useful in many use-cases in healthcare sector [7].Clustering of Patient data based on various factors such as BMI & Average Glucose Levels. High BMI is associated with risk of Obesity, Type 2 Diabetes, Cardiovascular Diseases, Joint Problems whereas High Glucose levels associated with risk of Diabetes, Neuropathy, Retinopathy, Cardiovascular problems and so on. Medical Image Segmentation, it plays a critical role in healthcare sector by extracting relevant information from the images and assisting healthcare professionals in making accurate diagnoses and treatment decisions, some of the important aspects of image segmentation are.

- It helps in identifying, characterize and early detection of the region of interests such as Tumors, abnormal structures in MRI, CT, X-Rays.
- It helps in quantification of various features such as, the tumor volume, tissue characteristics and so on.
- Segmented medical images serve as the basis of developing virtual and augmented reality models.
- Helps in personalized medication.
- Plays an important role in brain mapping and neuroimaging studies, helps in identification and localization of specific brain structures.

K-means clustering can also help in various other use-cases significantly such as patterning the length of stay in hospitals, it helps in understanding patient characteristics and differentiating disease subtypes based on their hospital durations. Clustering can help stratify patients based on their medical profiles, symptoms, or disease progression, allowing for personalized treatment plans and improving patient outcomes. Clustering can analyze large datasets related to public health, helping to identify high-risk regions for certain diseases, patterns of disease spread, and potential preventive measures. K-means clustering can assist in optimizing the allocation of medical resources by identifying areas with higher healthcare demand or specific medical needsand alsocan analyze continuous patient monitoring data, such as vital signs, to identify patterns that may be indicative of deteriorating health or early warning signs of certain medical conditions.

### 3.1 ClusteringofPatient data basedonfeatures.

This use case showcases how K-Means clustering can be applied to patient data, considering BMI (Body Mass Index) and average glucose level as features. Applying K-Means clustering to patient data on these features can bring several advantages to healthcare providers.

**International Journal of Management, IT & Engineering**
Vol. 13 Issue 08, August 2023
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

**Patient Segmentation**: K-Means clustering can help healthcare providers to group patients with similar characteristics based on features like BMI and Average Glucose levels.
This can help in finding patient subgroups which may have distinct health needs, risk factors and treatment process.

**Personalized Treatment Plans:** From the identified clusters of patients, healthcare providers can create treatment plans to directly address the specific need of each patient group, personalized approach can lead to better patient outcomes.

**Early Detection and Health Trends:** In case of analyzing public health data, K-Means can help significantly in identifying the health trends and the early detection of diseases that can lead to better disease management.

**Resource Allocation**:Patient clusters can help healthcare providers to allocate their resources more efficiently. They can prioritize resources, such as staff (nurse, doctors), equipment, and facilities, to target the patient clusters that require more attention or have higher health risks.

**Research and Clinical Trials**: K-Means clustering can also be used in research to identify similar groups of patients for clinical trials or studies. Researchers can focus on specific patient clusters to gain more accurate and applicable insights.

**Patient Education and Engagement:** Patient clusters can be used to develop targeted educational materials and interventions to improve patient understanding of their health conditions and encourage them to be more engaged in managing their health.

K-Means clustering offers healthcare providers a powerful tool for gaining valuable insights from patient data, leading to improved patient care, resource optimization, and better decision-making in healthcare delivery. Applying K-Means to the dataset [8] based on BMI and Average Glucose Levels is very common and a straightforward process with following steps needs to be applied.

**Data Preprocessing:** Ensure the dataset is cleaned and any missing or erroneous data points are handled appropriately like handling of null or NaN values in columns. Then out of the dataset get the features needed for clustering BMI & Average Glucose Levels, in the features we need to check the values and need to scale or normalize the features if needed because they must have a similar scale else one feature can dominate the other feature.

**Choosing the Number of Clusters (K) and Elbow method:** The elbow method is primarily used when applying K-Means clustering to determine the best value of K, so that the number of clusters would produce meaningful and distinct groups. Elbow method is always recommended first to apply in case of limited domain expertise and this method can helps us to arrive at an informed decision. It's important to note that the elbow method provides a heuristic for selecting K and may not always have a clear, definitive elbow point. The choice of K can sometimes be subjective based on the specific context of the data and the problem at hand.

**Applying K-Means:** K-Means can be implemented using Python, Pandas, scikit-learn library, we need to select the number of clusters (K) and fit the K-Means model to the selected features such as BMI and Average glucose levels.

**Interpreting the results:** The results can be analyzed on the clustered data, access the cluster assignments for each data point, as well as the centroids. Explore and visualize the clusters to gain insights into the characteristics of each group.

**Evaluating the Clustering (Optional):**Though K-Means doesn't have an intrinsic evaluation metric, external validation measures like Adjusted Rand Index or silhouette score should be utilized.Applying K-Means on individual columns can provide several benefits and insights that might be useful in understanding the distribution and characteristics of the data within each feature like Data Exploration, Feature Understanding and effective feature processing, Identifying Anomalies, can help in validation once we run the algorithm with multiple features

**International Journal of Management, IT & Engineering**
Vol. 13 Issue 08, August 2023
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

and feature comparison. On execution of K-Means on the dataset with only BMI feature with number of clusters 3 categorize the data as,

```
Cluster 1: min = 36.4, max = 97.6
Cluster 0: min = 26.0, max = 36.3
Cluster 2: min = 10.3, max = 25.9
```

Utilizing visualization can enhance the understanding of data insights regarding the three clusters. Cluster two exhibits the lowest values, while Cluster zero displays moderate values, and Cluster one falls within the highest danger zone. Visualizing these clusters can helpin clearer and more comprehensive understanding of the data distribution and the varying degrees of the data within each cluster.

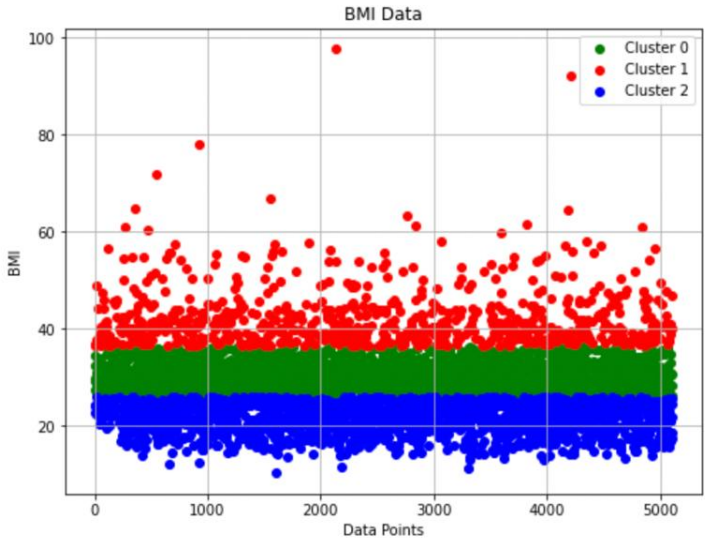


Figure 3. *BMI Visualization*

On applying K-Means on individual column helps in effective feature comparisons also it helps in finding whether the data is scaled appropriately or not. Then applying K-Means on more than one features BMI & Average glucose level on the given dataset produces following results.

```
Cluster 1: min bmi = 14.2, max bmi = 71.9, min glucose = 162.72, max glucose = 271.74
Cluster 2: min bmi = 10.3, max bmi = 61.6, min glucose = 97.06, max glucose = 162.93
Cluster 0: min bmi = 11.3, max bmi = 97.6, min glucose = 55.12, max glucose = 97.14
```

Utilizing visualization can enhance the understanding of data insights regarding the three clusters.

**International Journal of Management, IT & Engineering**
Vol. 13 Issue 08, August 2023
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

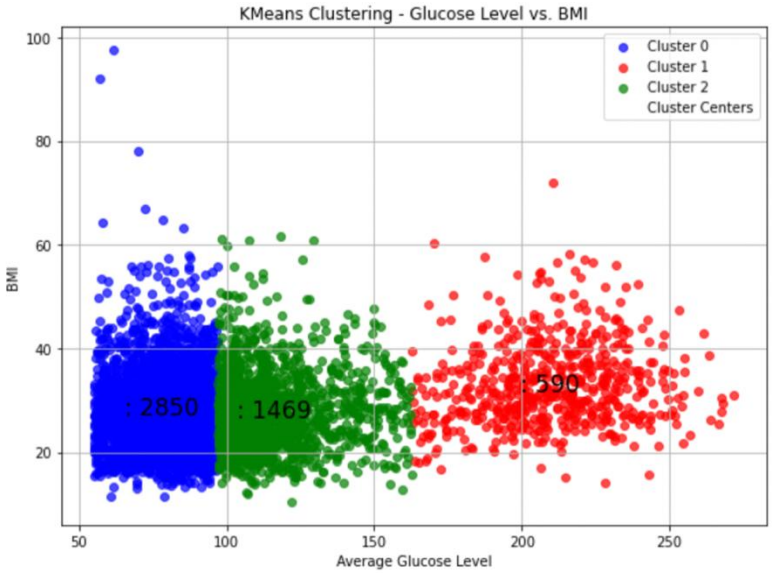


Figure 4. *BMI & Avg Glucose Level Visualization*

Applying K-Means clustering on BMI and Average Glucose level revealed distinct patient groups based on these two features. The clustering helped identify individuals with similar BMI and glucose level patterns, enabling personalized treatment plans and early detection of potential health risks. This data-driven approach provides valuable insights for healthcare providers to optimize resource allocation and enhance patient care specially in more complex disease dataset.

**3.2MedicalImageSegmentationusing K-means.**

Image segmentation is the process of partitioning an image into multiple segments, segmenting an image changes the representation of an image into something that is more meaningful and easier to analyze. Processing an entire image is not useful because many parts of the image may not contain any useful information, therefore by segmenting an image we can make use of only the important segments for processing. An image is a set of pixels, in image segmentation, pixels that have similar attributes are grouped together. In healthcare, image segmentations play a crucial role in various aspects in medical imaging analysis and diagnosis.

**Tumor detection & Quantitative Analysis**: Image segmentation is widely used to identify the tumors in MRIs, CT scans. By segmenting tumors helps in assessing the size, shape and location of the tumor that is essential for diagnosis, treatment planning and monitoring disease progression.

**Drug Discovery:**Image segmentation can be used to study the effects of drugs on tissues, example is, it can be used to segment the liver in an MRI scan, and then track how the liver changes over time after patient takes a drug. This information can also be helpful in developing new drugs.

**Diabetic retinopathy screening:** Image segmentation is being used for diabetic retinopathy, a major cause of blindness in people with diabetes.

**Cardiac disease diagnosis:** Image segmentation is being used to diagnose heart related diseases such as heart failure, coronary artery disease. This can help doctors to assess the severity of the disease and to plan treatment.

**Brain surgery planning:** Image segmentation is being used to plan brain surgery. This can help healthcare providers to visualize the brain and to identify the structures that need to be preserved during surgery.

These are just a few examples of how image segmentation is being used in healthcare today. As the technology continues to improve, we can expect to see even more applications for image segmentation in the future.

Applying K-Means on images for segmentation purposes offers many advantages, making it preferred choice for segmentation it is relatively simple and fast clustering algorithm, making it efficient for image segmentation especially for large datasets and it is scalable and can handle many data points (pixels in the case of images). This is essential for processing high-resolution images commonly encountered in medical imaging and satellite imagery.

The centroids obtained during the K-Means process can serve as representative colors or features of the segmented regions. This can be useful in cases where compact representations of segmented regions are needed.K-Means produces easy-to-understand results as each pixel is assigned to a specific cluster (segment). This makes it straightforward to interpret and analyze the segmentation outcomes.The K-Means algorithm iteratively updates cluster assignments and centroids to minimize the within-cluster sum of squares. This process typically leads to convergence and reasonably good segmentation results and the segmented regions obtained can serve as a useful preprocessing step for other image analysis algorithms, such as object recognition, or further refinement using more sophisticated segmentation techniques.
Applying K-Means to the dataset [9] of Brain MRI image segmentation following general steps needs to be applied. Here's a general outline of how to use K-Means for image segmentation:

**Read Image and convert it to Grayscale**: Read the image from a specified file path and convert the RGB image to grayscale. The NumPy libraries mean function is applied on the image on 3$^{rd}$ axis to compute the mean value of every pixel that results in a 2D grayscale image.

**Reshape the image:** To apply K-means algorithm, the 2D grayscale image needs to be converted to 2D NumPy array, the reshape function should be utilized for the same.

**Choose the Number of Clusters (K) and Initialize the Clustering algorithm:**The K-means clustering algorithm is initialized with the specified number of clusters say k=6, that means the algorithm will try to partition the image into 6 segments based on pixel intensity similarity.

**Fit the K-means Model to the Image Data:**The K-means algorithm is applied to the and it attempts to group similar pixels into the specified number of clusters say 6. The algorithm iteratively updates the cluster centers to minimize the sum of squared distances between each pixel and its assigned cluster center. After fitting the data, the K-means algorithm converges, and the cluster centers become fixed.

**Get the cluster label for each pixel:** After the algorithm completion, it assigns a cluster label to each pixel in the image.

**Create a Segmented Image:**The cluster labels retrieved from K-means are reshaped to match the original shape of the grayscale image, that is the segmented image. Each pixel in the segmented image represents the cluster to which that pixel has been assigned.

To identify the clusters assigned for every pixel the 'labels_' attribute should be utilized, this attribute contains the cluster assignments for each data point meaning each pixel in the image.

**International Journal of Management, IT & Engineering**
Vol. 13 Issue 08, August 2023
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

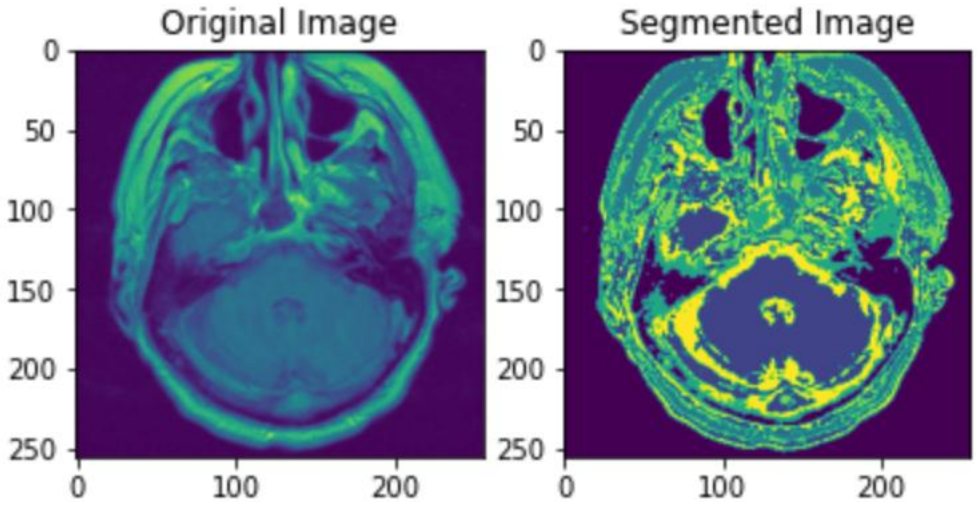


Figure 5. *Image Segmentation*

K-Means clustering for image segmentation has various advantages in healthcare sector some of them the main ones are Tumor detection, tissue segmentation and it's a simple algorithm for image data as well.

**4. Future Scope**

AutoML on Cloud [10][11] is a collection of powerful machine learning (ML) services that helps in building and deploying Machine Learning models without having to be an expert in Machine Learning. It uses a various technique, including K-means clustering, to automatically find the best ML model for the data.

AutoML on Cloud can use K-means clustering to automatically find the best clustering results for the data. It will typically try several different values for the number of clusters (K) and then select the value that produces the best clustering results. AutoML will also typically try different initialization methods and stopping criteria to improve the performance of the K-means algorithm.

To utilize K-Means clustering in Google Cloud Platform's AutoML, upload the dataset, select the algorithm in the AutoML wizard, specify the number of clusters (k), choose initialization/stopping criteria, and initiate the clustering job, an entire pipeline can be created to automate the process [12]

**5. ResultsandAnalysis**

In the first use-case, K-means clustering algorithm is applied to patient data on features like BMI and Avg Glucose level, through the algorithm we aimed to identify distinct patient clusters based on BMI and Glucose data features to uncover potential subgroups with similar characterisitics.

Upon applying the K-means clustering algorithm, the algorithm is effectively able to group patients into disitinct clusters, each cluster represents a group of patients with similar attributes. This segmentation can help provide effective and personalized treatment to the patients in every cluster.

In the second use-case, K-means algorithm is used for image segmentation, we utilized MRI images to segment these images into distinct regions of interest. Through visual inspection we can clearly see that K-means is successfully able to partition the medical images into relevant segments.

## 6. Conclusion

The paper explained what K-means clustering is and how it works, detailing its fundamental principles. K-means is an unsupervised learning algorithm used to partition data into clusters based on similarities, making it a valuable tool for various applications, including the healthcare sector. By utilizing K-means clustering, healthcare professionals can find valuable insights as such as clustering of patients based on features and medical image segmentation. Algorithm is fairly easy to apply though it's iterative has various advantages like identification of Tumors, Tissue segmentation, proven identification of diabetic retinopathy and so on. The future direction of the algorithm is to utilize the AutoML feature provided by almost every cloud provider, AutoML streamlines the process of model selection and hyperparameter tuning enabling more accessible and efficient usage of K-means algorithm. As the healthcare industry continues to embrace the potential of artificial intelligence and machine learning, leveraging algorithms like k-means through AutoML in platforms such as GCP can lead to transformative improvements in healthcare practices and outcomes.

## References

[1] Pham DT, Dimov SS, Nguyen CD. Selection of K in K-means clustering. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science. 2005;219(1):103-119. doi:10.1243/095440605X8298
[2] Okada, S., Ohzeki, M. & Taguchi, S. Efficient partition of integer optimization problems with one-hot encoding. Sci Rep 9, 13036 (2019). https://doi.org/10.1038/s41598-019-49539-6.
[3] Liu, F., & Deng, Y. (2020). Determine the number of unknown targets in open world based on elbow method. IEEE Transactions on Fuzzy Systems, 29(5), 986-995.
[4] Li, X., Zhang, P., & Zhu, G. (2019). DBSCAN clustering algorithms for non-uniform density data and its application in urban rail passenger aggregation distribution. Energies, 12(19), 3722.
[5] Haraty RA, Dimishkieh M, Masud M. An Enhanced k-Means Clustering Algorithm for Pattern Discovery in Healthcare Data. International Journal of Distributed Sensor Networks. 2015;11(6). doi:10.1155/2015/615740
[6] Spini, G., van Heesch, M., Veugen, T., & Chatterjea, S. (2019). Private Hospital Workflow Optimization via Secure k-Means Clustering. Journal of medical systems, 44(1), 8. https://doi.org/10.1007/s10916-019-1473-4
[7] Armstrong, J. J., Zhu, M., Hirdes, J. P., & Stolee, P. (2012). K-means cluster analysis of rehabilitation service users in the home health care system of Ontario: Examining the heterogeneity of a complex geriatric population. Archives of physical medicine and rehabilitation, 93(12), 2198-2205.
[8] https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
[9] https://www.kaggle.com/code/hossamfakher/mri-segmentation-classification
[10] Kunduru, Arjun Reddy. (2023). Effective Usage of Artificial Intelligence in Enterprise Resource Planning Applications. International Journal of Computer Trends and Technology. 71. 73-80. 10.14445/22312803/IJCTT-V71I4P109.
[11] Kunduru, Arjun Reddy. (2023). Cloud Appian BPM (Business Process Management) Usage In health care Industry. IJARCCE. 12. 339-343. 10.17148/IJARCCE.2023.12658.
[12] Sameer Shukla, (2022). Developing Pragmatic Data Pipelines using Apache Airflow on Google Cloud Platform. *International Journal of Computer Sciences and Engineering*, 10(8), 1-8.