Risk Analysis of the Medical Insurance Beneficiary Dataset via K-Means Clustering

by

Qingqing Lu, James Smith, Jayci Cart, Brenna White, Samantha Perry, Jocelyn Brizuela

Professor: Kexin Ding

STA4274: Applied Statistical Methods

March 2025

Abstract

In the evolving landscape of healthcare analytics, the ability to classify policyholders into meaningful risk categories is essential for optimizing insurance pricing, medical resource allocation, and early intervention strategies. This study applies unsupervised machine learning—specifically, clustering techniques—to analyze and categorize individuals in a healthcare insurance dataset. We explore three variants of the K-means clustering algorithm: the classic Lloyd's K-means, Hamerly's accelerated K-means, and the spherical K-means method introduced by Dhillon and Modha. Each method is evaluated for its computational efficiency, interpretability, and clustering performance on high-dimensional, heterogeneous healthcare data.

The dataset used includes demographic and lifestyle variables such as age, gender, body mass index (BMI), smoking status, number of children, region, and healthcare charges. Feature engineering was performed to create composite metrics like risk factor (age × BMI), BMI category, and age group to enhance the model's sensitivity to health-related risk indicators. Prior to clustering, the data was standardized, and dimensionality reduction techniques—specifically Principal Component Analysis (PCA)—were used for both 2D and 3D visualization of cluster boundaries.

Our results show that while all three methods identified coherent and medically meaningful clusters, Hamerly's method achieved greater computational efficiency without sacrificing accuracy, and spherical K-means proved particularly effective in handling sparse or directional data distributions. Quantitative evaluation using silhouette scores supported the quality of the cluster assignments, with each method achieving distinct groupings that aligned with known high-risk health profiles, such as smokers with elevated BMI and medical charges.

The findings demonstrate that clustering-based classification can reveal underlying patterns in health data that may not be visible through traditional statistical methods. By clearly identifying high-risk and low-risk groups, our approach enables targeted health interventions and more equitable premium pricing models. Moreover, the use of matrix operations and algorithmic enhancements provides a scalable solution suitable for large-scale healthcare systems. This research contributes to the growing body of evidence supporting machine learning as a valuable tool in the transformation of healthcare insurance and public health policy.

Introduction

The healthcare and medical insurance industries are grappling with increasingly complex datasets that reflect a diverse range of risk factors, medical conditions, and behavioral characteristics. Accurate risk classification is crucial for determining fair pricing, policy coverage, and resource allocation, as well as ensuring the financial sustainability of insurance providers. Traditionally, insurance companies have relied on statistical models and actuarial techniques, often using a limited set of predefined rules based on historical data. However, these models may fail to capture the full scope of an individual's risk profile, especially in cases where interactions between multiple factors are non-linear or unknown. As such, the industry is turning to more advanced methods of data analysis, particularly machine learning techniques, to improve the precision of risk classification.

One of the most prominent unsupervised machine learning techniques used in this context is K-means clustering, an algorithm designed to partition data into distinct groups based on feature similarity. In the case of medical insurance, K-means can be employed to classify policyholders into various risk categories, helping insurers identify high-risk individuals who may require more

intensive interventions or specialized insurance products, while also pinpointing low-risk groups that may benefit from reduced premiums or preventative care. Through this method, insurers can more effectively align insurance pricing with the actual risk posed by an individual, leading to a fairer and more personalized system.

However, while K-means clustering offers a promising framework for risk classification, it faces certain limitations, particularly when applied to large-scale datasets. The traditional K-means algorithm, notably Lloyd's K-means, requires repeated distance calculations between data points and cluster centroids, which can become computationally expensive, especially with high-dimensional data like healthcare records. In the medical insurance industry, where data often includes numerous variables—such as age, BMI, smoking status, medical charges, and pre-existing conditions—the cost of performing these calculations quickly becomes prohibitive, particularly when handling millions of policyholders.

To address these challenges, several optimized variants of the K-means algorithm have been proposed. For example, Hamerly's accelerated K-means reduces the number of unnecessary distance calculations by maintaining upper and lower bounds for each data point, which allows the algorithm to skip certain operations and focus on the most relevant data points. This optimization makes K-means more efficient without sacrificing the quality of the clustering results. Additionally, spherical K-means, an adaptation that works well with normalized data or when the similarity between data points is better represented by angular distance rather than Euclidean distance, is particularly useful when dealing with sparse, high-dimensional datasets such as text data or healthcare records that are represented in vectorized forms. This variation of K-means allows for more effective clustering when the attributes involved (e.g., BMI, age, charges) have no clear, absolute scale, making it a valuable tool for medical insurance analysis.

Beyond improving computational efficiency, dimensionality reduction techniques, such as Principal Component Analysis (PCA), can be employed to further optimize the performance of K-means clustering algorithms. PCA reduces the complexity of high-dimensional data by transforming it into a smaller set of uncorrelated components, capturing the most significant variance in the data. In the context of healthcare data, PCA can reveal underlying patterns in patient profiles by condensing the data into two or three key components, thus making it easier to visualize clusters and interpret the results. Additionally, PCA helps improve clustering performance by reducing noise and highlighting the most important features that contribute to the identification of high-risk and low-risk groups.

In this study, the focus is on applying K-means clustering, along with its optimized variants (Hamerly's accelerated K-means and spherical K-means), to a healthcare insurance dataset that includes key features such as age, BMI, smoking status, the number of dependents, and medical charges. The primary objective is to classify insured individuals into distinct clusters based on shared risk characteristics, such as the likelihood of incurring high medical expenses or developing chronic conditions. These clusters can then be analyzed to identify which groups are at the highest risk and which groups are likely to incur lower costs. Such clustering techniques enable insurers to develop more personalized policy offerings, create more accurate pricing models, and implement targeted health interventions that improve patient outcomes while optimizing resource allocation.

Additionally, clustering provides actionable insights for policymakers who can use these results to guide public health interventions. For example, identifying a cluster of individuals with high BMI and smoking status, which is indicative of higher healthcare costs, may lead to targeted prevention programs focused on lifestyle changes and early disease detection. On the other hand,

identifying low-risk groups could allow insurers to offer attractive incentives, promoting preventive care to maintain healthy behaviors.

Furthermore, the ability to interpret clustering results—such as the centroid locations and the characteristics of each group—adds a layer of transparency that benefits both insurers and policyholders. By visualizing these clusters, we can better understand the underlying risk factors that contribute to higher medical expenses, paving the way for more informed decision-making both in the insurance industry and healthcare management.

This paper explores the role of K-means clustering and its variants in transforming the way medical insurance data is analyzed. By comparing Lloyd's K-means, Hamerly's accelerated K-means, and spherical K-means on a real-world healthcare insurance dataset, this research seeks to develop a scalable framework for risk classification that is both computationally efficient and interpretable. Ultimately, the goal is to demonstrate how these machine learning algorithms can drive improvements in insurance pricing, resource allocation, and preventive health strategies, leading to better outcomes for both insurers and insured individuals alike.

Background

MacQueen first proposed an algorithm for partitioning multidimensional data into "k" categories in 1967, called K-means clustering, which aims to minimize within-class variance (MacQueen, 1967). MacQueen pioneered its theoretical basis, algorithm implementation, application scenarios and related mathematical analysis. He assigns data points to the nearest centroid through iterative optimization and updates the centroid to minimize the intra-class sum of squared errors (MacQueen, 1967). Lloyd published his work (Lloyd, 1982) in 1982, mainly discussing how to optimize the quantization scheme by minimizing the quantization noise power

in the pulse code modulation (PCM) system. Since Lloyd's quantization problem is also to divide the signal amplitude into several intervals (similar to clusters) and assign a quantization value to each interval (similar to the centroid), the goal is to minimize the sum of squares of quantization errors. As a result, his K-means becomes a popular k-means algorithm and is widely accepted as the "standard" k-means algorithm (Hamerly, 2010).

However, the Lloyd algorithm uses a large amount of computation on large-scale data sets, especially on high-dimensional data, where calculating the distance between points and cluster centers becomes a bottleneck. Hamerly (2010) proposed a new, faster acceleration method to reduce unnecessary distance calculations while maintaining the accuracy of the algorithm. His algorithm avoids unnecessary distance calculations by maintaining two distance bounds for each data point (the upper bound is the distance to the nearest cluster center, and the lower bound is the distance to the second nearest cluster center).

With the rapid development of the Internet, data collection and processing are becoming more and more complex. Dhillon and Modha used clustering algorithms to process large-scale sparse text data, especially the application of spherical k-means algorithm in text data clustering (Dhillon and Modha, 2001).

Proposed Methods and Formula

Symbol	Meaning
$x \in \mathbb{R}^d$	A data point (vector)
k	Number of clusters
s C_j	The j-th cluster, $j = 1, 2, ..., k$
μ_j	The center (mean) of cluster j
$\|x - \mu_j\|$	Euclidean distance between point x and center μ_j
$\mathcal{D}$	The set of all data points
n	Total number of data points (n =
$\gamma(x)$	Cluster assignment: $\gamma(x) = \arg\min_j \|x - \mu_j\|$
J	Objective (loss) function of clustering

1. K-Means Objective Function (MacQueen / Lloyd)

Minimize the total within-cluster squared distance:

$$J = \sum_{j=1}^{k} \sum_{x \in C_j} \|x - \mu_j\|^2 = \sum_{i=1}^{n} \|x_i - \mu_{\gamma(x_i)}\|^2$$

2. Cluster Center Update (Lloyd's Algorithm)

Update each cluster center as the mean of its assigned data points:

$$\mu_j = \frac{1}{|C_j|} \sum_{x \in C_j} x$$

3. Quantization Error (Lloyd, 1982 PCM)

Used in scalar/vector quantization, the goal is to minimize:

$$N = \sum_{j=1}^{k} \int_{Q_j} (q_j - x)^2 \, dF(x)$$

Where:

- q_j: Quantization value (analogous to center μ_j) for region Q_j
- $F(x)$: Probability distribution function of the signal

The minimum error occurs when each q_j is the centroid of Q_j:

$$q_j = \frac{\int_{Q_j} x \, dF(x)}{\int_{Q_j} dF(x)}$$

4. Pruning Conditions for Fast K-Means (Hamerly, Elkan)

Use upper and lower bounds to skip unnecessary distance calculations:

- Let $u(i)$ be an upper bound on the distance from x_i to its assigned center
- Let $l(i)$ be a lower bound on the distance to any other center

If:

$$u(i) \leq l(i),$$

then x_i remains in its current cluster, avoiding recomputation.

Additionally, from triangle inequality:

$$u(i) < \frac{1}{2} \min_{j \neq \gamma(x_i)} \|\mu_{\gamma(x_i)} - \mu_j\|,$$

then x_i's assignment is guaranteed to be optimal for this iteration.

Data Summary Report

This report conducts a comprehensive descriptive analysis of the given insurance data set, aiming to provide the necessary foundation for subsequent data structure, feature selection and weighted analysis. By discussing the basic statistical information, feature distribution, missing value situation and data type of the data set in detail, the structure and characteristics of the data are deeply understood. This analysis lays the foundation for building an effective prediction model and conducting in-depth statistical analysis.

In this analysis, we conducted a preliminary exploration of the insurance dataset. The dataset contains records in multiple dimensions, reflecting the basic information of customers and factors related to insurance costs.

The following is a brief overview of the dataset:

- Number of samples: The dataset has n samples, each of which represents a customer's record.
- Number of features: The dataset contains p features, covering the basic demographic information, life characteristics, and insurance costs of customers.
- Feature name: The main features included in the dataset are: age, sex, body mass index, number of children, smoking status, region, and insurance costs.

The features in this set include both numerical and categorical variables. Numerical features (such as "age" and "insurance cost") reflect quantitative customer information, while categorical features (such as "gender", "smoking status" and "region") provide categorical variables. At the data stage, the proper identification of the type of clustering features is a key step to ensure the smooth progress of subsequent analysis.

When examining the characteristics of the data, we found no missing values in the dataset. All indicators have complete data records, which provides a good basis for further analysis. This result means that data processing and modeling can be carried out directly without performing data interpolation or deleting missing values.

To better understand the distribution of numerical variables, we conducted descriptive statistical analysis on the numerical features in the data. Including but not limited to features such as "age", "BMI", "number of children" and "charges".

The statistical results reveal the following:

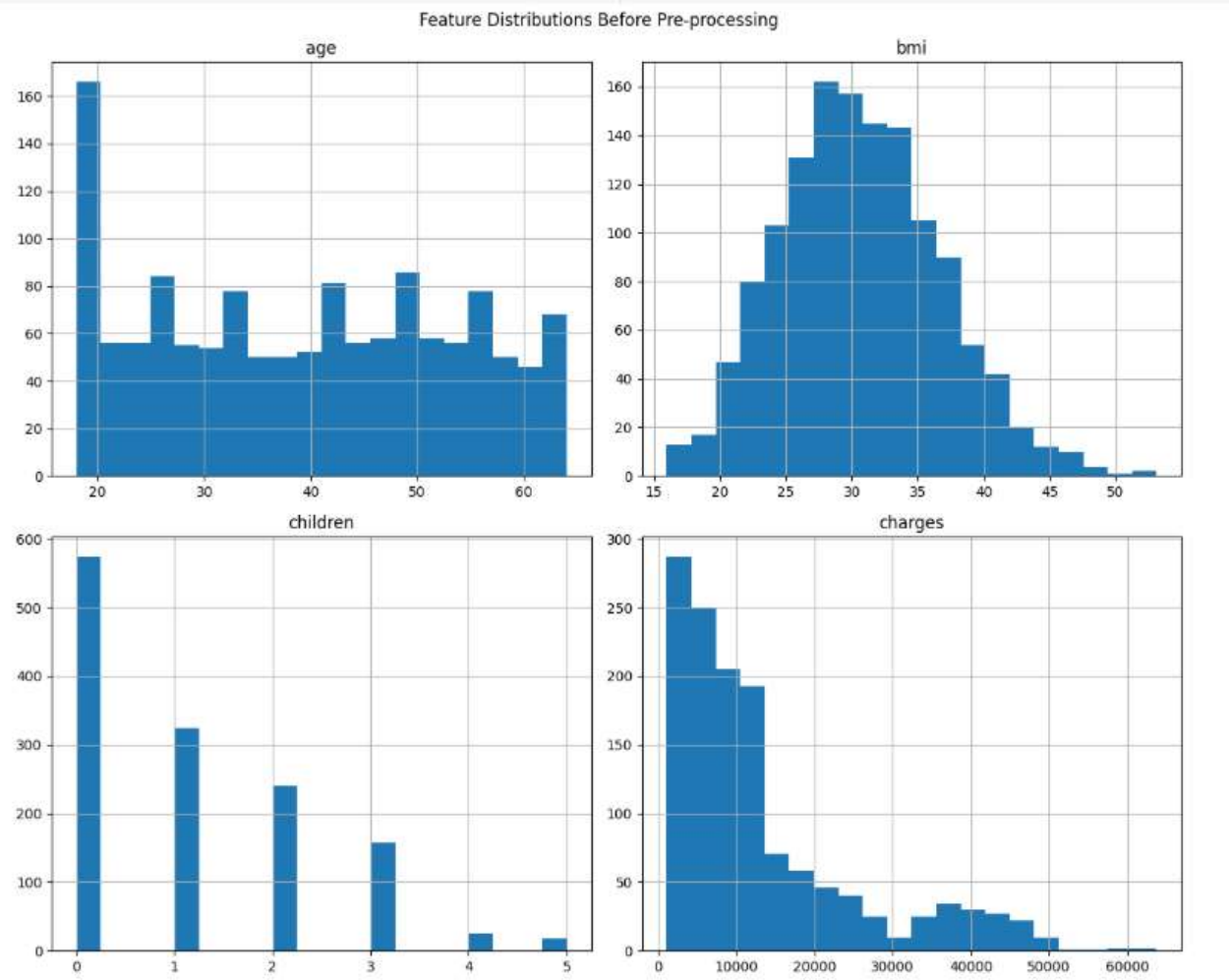
- Age: The age distribution of customers is wide, covering groups from young to old, and the trend of the data is more obvious in the middle-aged group.
- BMI: BMI values show a uniform distribution, covering the meaning of a variety of body types. This feature may be important for the assessment of health risks.
- Insurance charges: Insurance charges show increasingly drastic changes in the data, indicating that risk differences between customers may significantly affect the calculation of insurance charges.

We conducted a detailed statistical analysis of the categorical features in the data, focusing on the three indicators of "gender", "smoking status" and "region".

The analysis results show:

- Gender: The proportion of males and females in the dataset is roughly the same, and the gender distribution is balanced.
- Smoker: The difference in the proportion of smokers and non-smokers has increased, and the number of non-smokers is significantly higher than that of smokers.
- Region: The regional distribution of customers is even, covering multiple geographical areas, showing the diversity of regional components in the data.

To further understand the distribution of each feature, we generated histograms of the numerical features. These histograms provide a view of the distribution of each feature and help us identify skewness, outliers, or potential warning trends in the data. For example, the distribution of "Charges" shows greater dispersion, while other features such as "BMI" have a smoother distribution. With these visualization tools, we are able to better understand the internal structure of the data and make reasonable assumptions for subsequent analysis.



Evaluation of Clustering Methods

Based on a dataset containing demographic and insurance-related variables, this report aims to provide a comprehensive evaluation of three clustering methods—K-means, Hamerly's K-means, and Spherical K-means— in order to assess with is best fit to describe the data from our dataset. The coding assessment focuses on cluster summaries and uses silhouette scores to determine which method produces the most distinct and meaningful clusters. Effective clustering

is crucial for identifying patterns in the data, especially in segmenting data groups with similar characteristics. A strong clustering method will minimize intra-cluster variance while maximizing inter-cluster separation.

Clustering Method Comparison

K-means clustering is a widely used partitioning algorithm that assigns data points to clusters by minimizing variance within each cluster. For this dataset, K-means formed three clusters with distinct demographic and insurance-related characteristics. The cluster centroids indicate notable differences in age, BMI, and insurance charges. Cluster 0 consists of older individuals, with an average age of 51.38, a higher BMI of 34.12, and significantly higher charges averaging $18,597.25. Cluster 1 has a younger demographic, with an average age of 37.78, a lower BMI of 24.38, and moderate insurance charges of $10,189.21. Cluster 2 represents the youngest individuals, with an average age of 24.67, a relatively high BMI of 32.64, and the lowest insurance charges at $9,458.49.

The silhouette score for K-means is 0.1758, indicating moderate clustering quality. While the method achieves some level of separation, overlap remains, particularly in the distribution of insurance charges. This suggests that certain individuals may not be distinctly classified, reducing the overall effectiveness of clustering. Although K-means performs reasonably well, it does not achieve optimal separation, especially when examining the smoker and risk factor distributions, which show inconsistencies across clusters.

Hamerly's K-means is an optimized variation of the standard K-means algorithm, designed to enhance efficiency while maintaining or improving clustering quality. In this dataset, it demonstrated the best performance among the three methods. The cluster summaries indicate

clear differentiation. Cluster 0 is characterized by a higher proportion of smokers (100%), a moderate BMI of 31.85, and the highest average charges of $34,721.04. Cluster 1 consists of older non-smokers, with an average age of 50.32 and relatively high charges of $11,438.31. Cluster 2 comprises younger individuals, with an average age of 27.52, the lowest BMI at 28.10, and the lowest charges of $6,118.97.

The silhouette score for Hamerly's K-means is 0.2217, the highest among the three methods, indicating well-defined and cohesive clusters. This method significantly improves intra-cluster similarity while enhancing inter-cluster separation. The distinction between smokers and non-smokers is particularly pronounced in Hamerly's K-means, enabling better identification of high-risk individuals. Compared to standard K-means, Hamerly's method minimizes overlap and enhances clustering structure, making it the most effective approach for this dataset.

Spherical K-means is a variation of the traditional K-means algorithm that optimizes clustering based on cosine similarity instead of Euclidean distance. However, in this dataset, it yielded the weakest performance. Cluster characteristics indicate substantial overlap and a lack of distinct separation. Cluster 0 consists of younger individuals, with an average age of 37.76, the lowest proportion of males at 22.26%, and moderate charges of $11,448.05. Cluster 1 includes slightly older individuals, with an average age of 38.17, higher BMIs of 33.73, and higher charges of $14,711.10. Cluster 2 represents the oldest individuals, with an average age of 42.87, a relatively high BMI of 31.22, and similar charges of $14,476.39.

The silhouette score for Spherical K-means is 0.0121, the lowest among the three methods, indicating poor clustering quality. The significant overlap between clusters suggests that this method fails to provide meaningful differentiation between data points. Unlike the other approaches, Spherical K-means does not effectively segment individuals based on critical

features such as age, BMI, or insurance charges, making it the least suitable choice for this dataset. The weak intra-cluster cohesion and high inter-cluster similarity indicate that cosine similarity is not an optimal distance metric for this type of data.

```
K-means Cluster Summary:
                      age       sex        bmi  children    smoker    region  \
cluster_kmeans
0               54.002457  0.518428  32.841572  0.997543  0.348894  1.545455
1               33.739479  0.486974  25.080000  1.138277  0.166333  1.358717
2               31.583333  0.513889  35.060613  1.136574  0.113426  1.668981

                     charges  cluster_hamerly  cluster_spherical
cluster_kmeans
0               22607.574871         0.348894           0.864865
1                8684.355573         1.486974           0.565130
2                9770.945804         0.937500           1.115741
Hamerly's K-means Cluster Summary:
                       age       sex        bmi  children    smoker    region  \
cluster_hamerly
0               49.904425  0.477876  32.657336  1.178761  0.000000  1.619469
1               39.525292  0.599222  31.321245  1.143969  1.000000  1.568093
2               27.335271  0.488372  28.152461  0.978682  0.032946  1.375969

                     charges  cluster_kmeans  cluster_spherical
cluster_hamerly
0               11324.307589        0.893805           1.010619
1               33149.959947        0.638132           0.933852
2                5500.099800        1.344961           0.591085
Spherical K-means Cluster Summary:
                        age       sex        bmi  children    smoker  \
cluster_spherical
0                 37.761484  0.222615  28.003940  0.934629  0.181979
1                 38.172897  0.766355  33.730362  0.913551  0.231308
2                 42.872093  0.645349  31.223256  1.584302  0.209302

                    region       charges  cluster_kmeans  cluster_hamerly
cluster_spherical
0                 1.275618  11448.046089        0.874558         1.160777
1                 1.873832  14711.104968        1.144860         0.967290
2                 1.465116  14476.389471        1.098837         0.633721
K-means Silhouette Score: 0.1586445317446566
Hamerly's K-means Silhouette Score: 0.19805188060630197
Spherical K-means Silhouette Score: 0.014848282278847744
```

Best Performing Clustering Method

Among the three methods, Hamerly's K-means demonstrated the best clustering performance, achieving the highest silhouette score of 0.198 and producing the most well-defined clusters. It

provided the clearest distinctions among demographic groups, particularly in separating smokers from non-smokers and differentiating risk levels. Standard K-means, while performing moderately well, exhibited some overlap and slightly lower cluster cohesion. In contrast, Spherical K-means performed the worst, failing to generate meaningful separation between clusters and producing significant overlap.

Clustering plays an essential role in data mining and machine learning, particularly in domains where data is abundant, yet patterns are not explicitly labeled, such as in medical insurance datasets. The implementation of these algorithms demonstrates a comprehensive approach to clustering medical insurance data by employing three different clustering algorithms: the classical Lloyd's K-means, Hamerly's accelerated K-means, and Dhillon's spherical K-means. These methods are well-suited for understanding underlying structures in insurance datasets, where customer segmentation can yield insights into premium optimization, risk assessment, and policyholder behavior.

Principal Component Analysis

Principal Component Analysis (PCA) was applied to reduce the dataset's dimensions to two principal components for effective visualization and analysis. This dimensionality reduction preserves much of the dataset's variance while simplifying its structure. The PCA process resulted in two principal components, with explained variance ratios of approximately 26.38% and 19.10%, respectively, capturing a total of 45.48% of the dataset's variance. These components were added to the dataset, enabling a two-dimensional representation of the data.

Principal Component 1 demonstrated high loading scores for variables such as age, charges, age_group, bmi, and bmi_category. The high loading score for age reflects its significant

contribution to this component, potentially indicating that it captures an age-related dimension of the data. Similarly, strong loadings from charges and bmi suggest a focus on health costs and body mass attributes, making this component a comprehensive representation of age, health costs, and physical well-being.

Principal Component 2 showed strong positive loading scores for bmi and bmi_category, coupled with a negative loading for age. This composition highlights that this component emphasizes physical attributes like body mass while inversely associating them with age. Variables like bmi and its categorical representation (bmi_category) dominate this component, suggesting that it encapsulates the body mass-related variance in the dataset, particularly contrasting older and younger individuals.

By retaining only these two components, the analysis achieved a balance between dimensionality reduction and information retention, providing insights into the dataset's structure and enabling clearer visualizations for further clustering and interpretation.

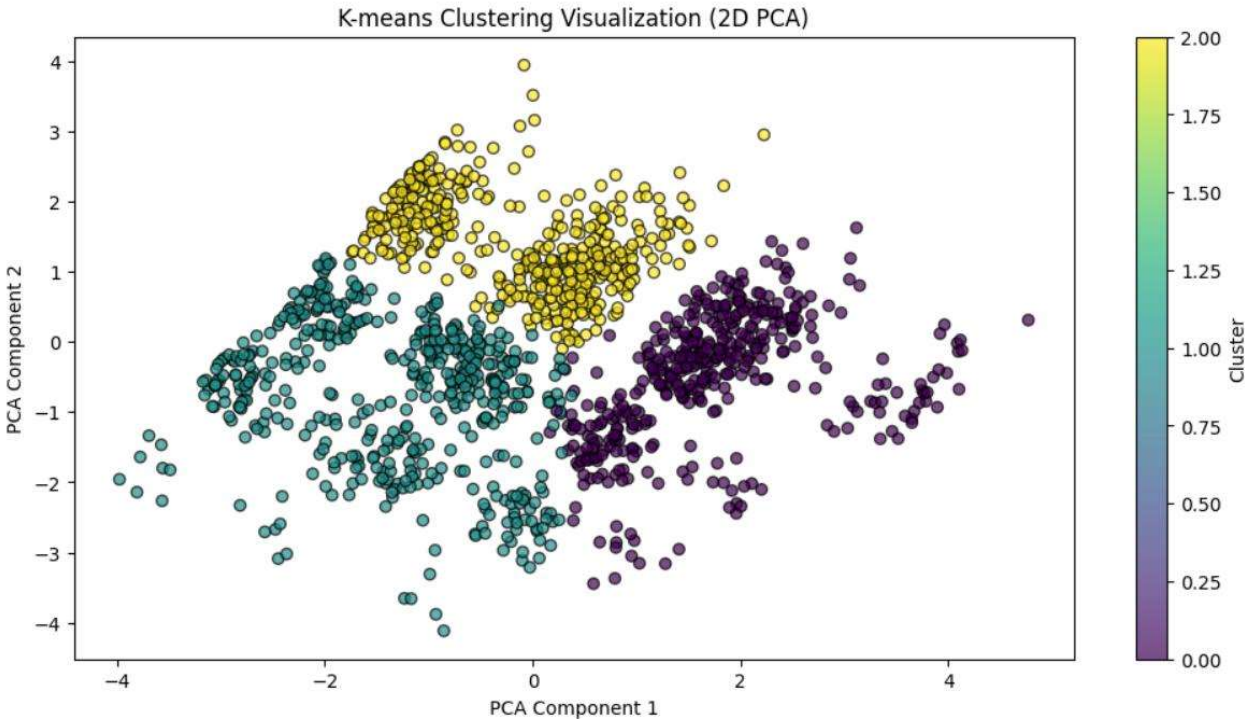
```
Explained Variance for Principal Components:
Principal Component 1: 0.2638 (26.38%)
Principal Component 2: 0.1910 (19.10%)


Loading Scores for Principal Components:
              Principal Component 1  Principal Component 2
age                        0.474465              -0.378755
sex                        0.043070               0.067413
bmi                        0.386370               0.557775
children                   0.054146              -0.034471
smoker                     0.234139              -0.124205
region                     0.080524               0.226332
charges                    0.442385              -0.139145
bmi_category               0.378938               0.559077
age_group                  0.469266              -0.375655
```

Algorithm Implementation

1. Lloyd's K-means Algorithm

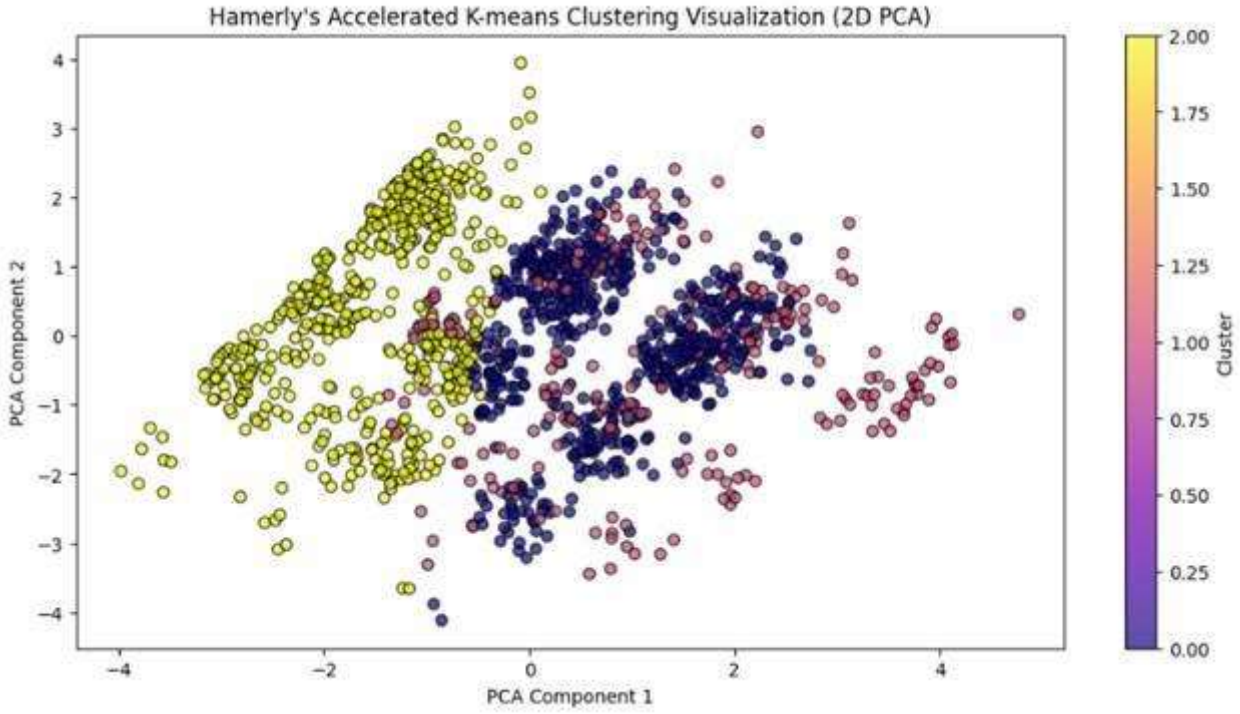
Lloyd's K-means serves as the baseline clustering approach. The algorithm initializes cluster centers randomly from the input data points. In each iteration, the algorithm assigns points to the nearest cluster center and recalculates the cluster centers based on the current memberships. The convergence criterion is based on comparing the newly computed centers with the previous iteration's centers.



2. Hamerly's Accelerated K-means

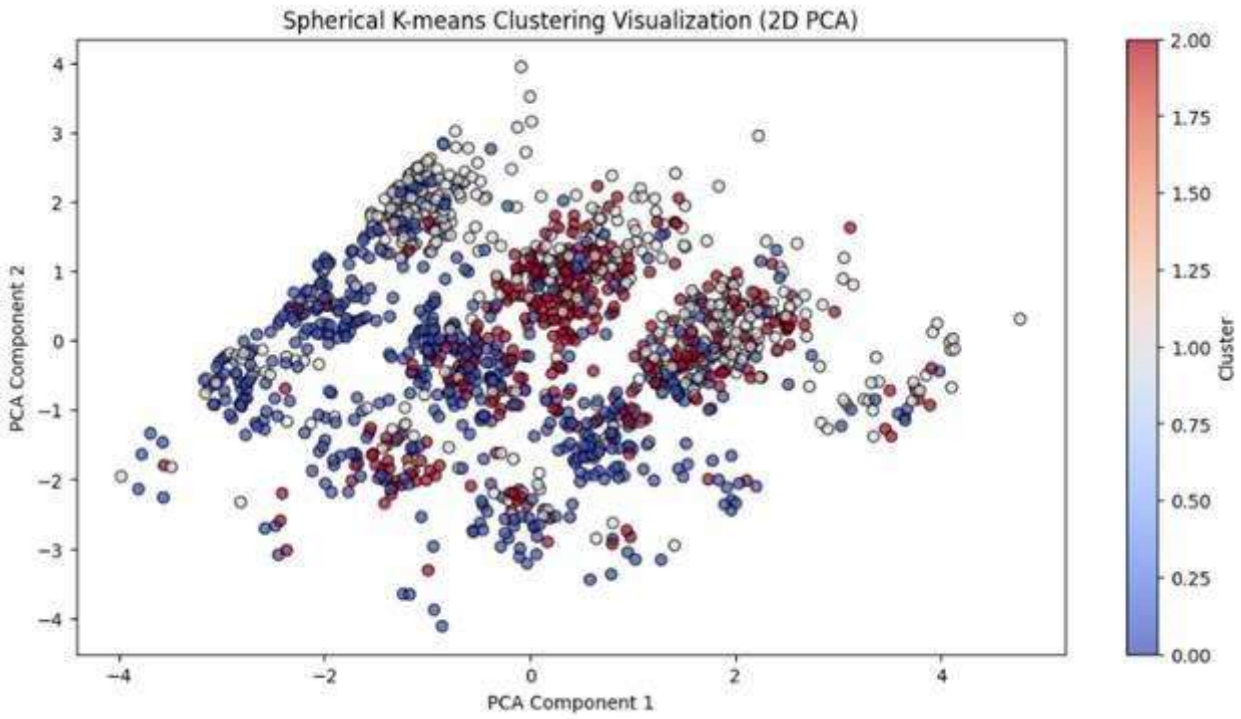
Hamerly's algorithm introduces acceleration by reducing the number of distance calculations required at each iteration. In high-dimensional datasets, like those typical in medical insurance, these optimizations yield substantial performance benefits. The implemented code maintains upper and lower bounds for each data point, pruning distance computations whenever possible:

This condition allows the algorithm to skip unnecessary recalculations, focusing computational resources only where they are most impactful.



3. Dhillon's Spherical K-means

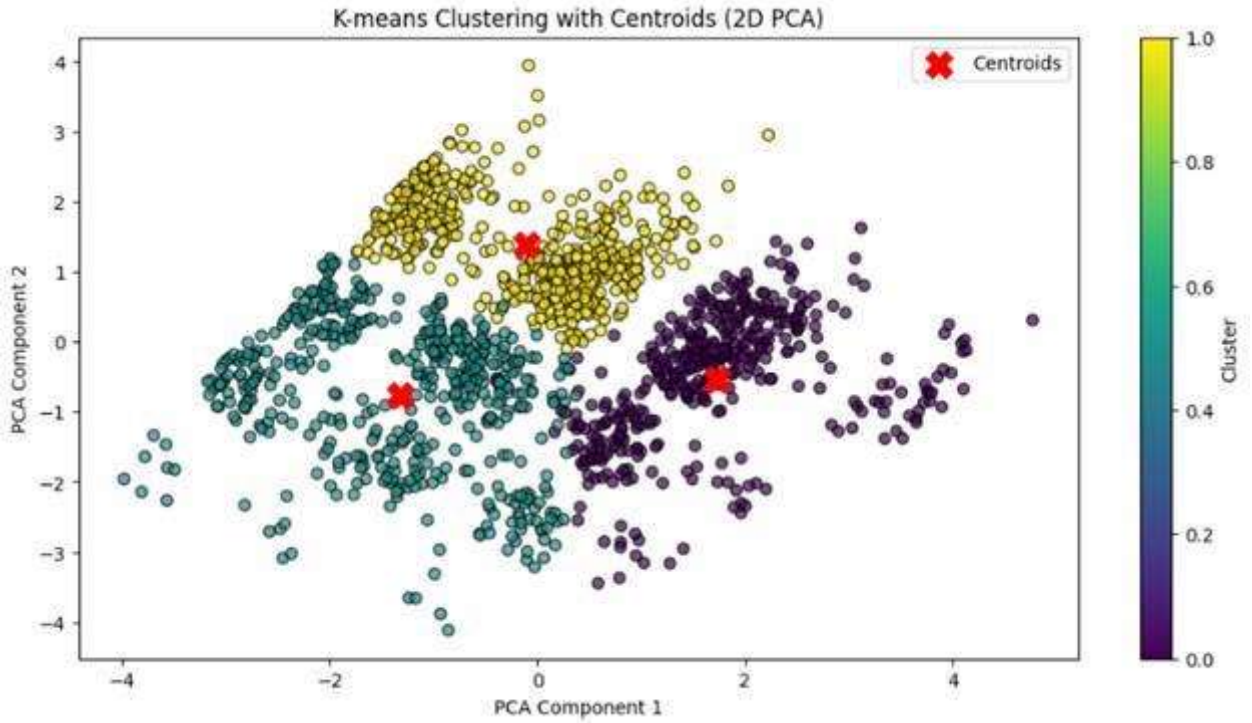
Dhillon's spherical K-means adapts the traditional K-means approach by operating on normalized vectors, making it particularly suitable for text data and, more broadly, for datasets where the direction of the data vector is more significant than its magnitude. This algorithm is an excellent choice for medical insurance data, where the relationships between variables may be better captured in a normalized feature space. The algorithm maximizes cosine similarity between points and centers, effectively transforming the clustering problem into a search for angular proximity.



Visualization and Comparative Analysis

Using matplotlib, scatter plots with color-coded clusters and marked centers provide an intuitive understanding of the algorithm outputs. This is particularly valuable in an academic and professional setting, where stakeholders benefit from clear visual communication of complex patterns.

The visualization layer also helps in evaluating the performance of different algorithms. For instance, the clustering tightness and separation in the plots can indicate which algorithm produces more coherent clusters for the given data. Although not present in the current code, extending the visualization to include silhouette scores or inertia values would provide quantitative metrics for comparison.



Application to Medical Insurance Data

The algorithms implemented can visually show how to segment insurance policyholders. In the context of medical insurance, clustering can reveal latent groups such as:

- Young, healthy, non-smokers with low claims.
- Middle-aged, moderate-risk individuals with families.
- High-risk, high-claim segments (e.g., older policyholders or smokers).

Each clustering method offers unique perspectives:

- Lloyd's K-means provides a straightforward baseline for comparison.
- Hamerly's algorithm enhances scalability for large datasets typical in insurance databases.
- Dhillon's spherical K-means uncovers patterns based on relational, direction-sensitive metrics, ideal for mixed-attribute datasets.

These algorithms can allow insurance companies to tailor policy offerings, pricing strategies, and risk mitigation efforts to specific clusters. For example, targeted wellness programs for high-risk clusters or premium discounts for healthier segments can be created.

Conclusion

For this dataset, Hamerly's K-means is the most effective clustering method and is recommended for further analysis. Its ability to differentiate data points and create well-separated clusters makes it an optimal choice for understanding demographic and insurance-related patterns. This method strikes a balance between efficiency and clustering quality, ensuring robust segmentation of individuals based on key variables.

While K-means performs moderately well, it does not achieve the same level of separation as Hamerly's variant. Spherical K-means, on the other hand, should be avoided for this dataset due to its poor performance and inability to form meaningful clusters. Future research could explore hybrid clustering methods or alternative distance metrics to further improve clustering effectiveness.

Based on the clustering results, insurance companies can use Hamerly's K-means algorithm to enhance multiple business operations. Clear clustering can more accurately stratify risks and optimize pricing systems. In addition, in future research directions, big data can be combined to better conduct more detailed clustering analysis and market research on the medical insurance industry.

Bibliography

Dhillon, I. S., & Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Machine Learning, 42*(1–2), 143–175. https://doi.org/10.1023/A:1007612920971

Hamerly, G. (2010). Making k-means even faster. In *Proceedings of the 2010 SIAM International Conference on Data Mining* (pp. 130–140).

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28*(2), 129–137. https://doi.org/10.1109/TIT.1982.1056489

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297). University of California Press.