Mesopotamian Journal of Artificial Intelligence in Healthcare
Vol.2025, **pp**. 187–196
DOI: https://doi.org/10.58496/MJAIH/2025/018 ; ISSN: 3005-365X
https://mesopotamian.press/journals/index.php/MJAIH

Research Article

Data Mining Driven Segmentation of Health Insurance Policyholders Using K-Means Clustering

Farah Ali Khairi[1,*], Laith Farhan[2] , , Oluwaseun A. Adelaja[3,]

[1] *Kufa Technical Institute, Al-Furat Al-Awsat Technical University, Kufa, Najaf, Iraq,*

[2] *School of Engineering, Manchester Metropolitan University, Manchester M1 5GD, UK,*

[3] *Information Communication and Technology Department, Lagos State University, Lagos 102101, Nigeria*

ARTICLE INFO

Article History

Received 05 May 2025
Revised 06 Jun 2025
Accepted 05 Jul 2025
Published 27 Jul 2025

Keywords

Data Mining

K-Means clustering

Health Insurance

Actuarial Analytics

Machine Learning

ABSTRACT

This study illustrates a data-driven approach to the segmentation of health insurance policyholders based on K-Means clustering of an open insurance dataset. Key demographic and financial features like age, body mass index (BMI), dependents, annual medical spending, and premium payment were normalized first to ensure comparability. The optimal number of clusters (k = 3) was determined using silhouette analysis, and three clusters were formed: (1) young, low-cost individuals, (2) middle-aged medium-cost individuals, and (3) old, high-cost individuals. Cluster centroids provide actionable profiles that can be utilized by insurers for target marketing, risk profiling, and development of customized plans. A set of visualizations scatter plots, boxplots, histograms, and bar charts illustrate the separation and within-distribution nature of these segments. The preprocessing workflow (missing value treatment, encoding of categorical features, and feature scaling) was encoded in a flowchart for reproducibility. Results demonstrate that straightforward-to-implement unsupervised learning techniques can yield interpretable customer segmentations, offering a foundation for more advanced predictive modeling and individualized insurance policies.

1. INTRODUCTION

Customer segmentation is a fundamental basis for strategic decision-making in the insurance industry, and it enables firms to tailor product provision, optimize risk assessment, and enhance customer satisfaction. In life and health insurance, accurate identification of homogeneous groups among insured customers enables tailored premium pricing and niche-focused marketing campaigns. In recent years, researchers have explored a number of unsupervised learning models to address this issue. For instance, Abdul-Rahman et al. used decision trees and K-modes clustering to segment life insurance clients, offering higher interpretability in categorical data environments [1]. Similarly, Gan and Valdez provided a comprehensive review of actuarial applications using clustering, emphasizing the significance of partition methods for risk stratification [9]. Traditional clustering techniques vary from distance-based algorithms such as K-Means and their fuzzy equivalent Fuzzy C-Means to model-based and spectral ones. The seminal paper by Hartigan put forward a general classification of clustering algorithms and laid the ground-level principles of partitioning data into well-separated, tight clusters [14]. Bezdek et al. later formalized the Fuzzy C-Means algorithm to enable mixed-membership clustering that can be used in real situations where observations might belong to multiple segments [3]. Alternatively, spectral clustering techniques such as Laplacian Eigenmaps map high-dimensional data into lower dimensions first before clustering in order to improve the separation of complex structures [2]. Hierarchical methods, as studied by Campo and Antonio, offer an alternative perspective using the formation of nested clusters that can identify multi-level risk factors in insurance [5].

Application-driven research has further advanced clustering methods for insurance analytics. Debener et al. combined supervised and unsupervised techniques to identify fraudulent claims, highlighting the pragmatic complementarity of classification and clustering in anomaly detection [6]. Hainaut's work on self-organizing maps illustrated how clustering with pattern recognition in neural networks can uncover concealed patterns in non-life insurance portfolios [12], while Hsu et al. illustrated the applicability of large-scale knowledge discovery based on self-organizing systems [15]. Complementary approaches, such as correspondence analysis and global similarity metrics, have also been proposed by Greenacre and Gower, respectively, to quantify relationships between mixed data types common in insurance data sets [11], [10]. Burt's

*Corresponding author. Email: farah.kairy.iku@atu.ed.iq

original factorial analysis of qualitative data also provided early theoretical grounds for decomposing categorical variables into continuous latent factors suitable for clustering [4].

Despite the expanse of clustering research, the K-Means algorithm ranks among the most used techniques due to its parsimony, scalability, and comprehensibility. By straightforward minimization of within-cluster variance, K-Means partitions policyholder data into distinct segments that easily profile and react to them. Dunn's index and the silhouette coefficient have been derived as best estimates of the number of clusters [7]. Additionally, K-Means extensions and improvements—such as those presented in Hainaut and Thomas's working note are available, each providing its own enhancements specific to insurance analytics, like initialization methods and cluster validity measures [13]. Fuzzy extensions and combination strategies have also been investigated to overcome the hard-assignment restriction of K-Means [8], [14].

In this paper, we apply K-Means clustering on a public health insurance dataset to uncover three major policyholder segments. The dataset includes demographic characteristics (age, gender, dependents), health characteristics (body mass index), and financial characteristics (annual expenses, premiums). A robust preprocessing pipeline encompassing missing-value management, categorical encoding, and numerical scaling is constructed to prepare data for clustering. We conduct silhouette analysis on identifying the optimal cluster number and examine cluster centroids to characterize each segment. Scatter plots, boxplots, and distribution histograms enable an intuitive interpretation of the clusters' properties. Our findings corroborate previous findings that simple unsupervised techniques can yield useful insights for focused insurance approaches [1], [9].

2. RELATED WORK

Clustering remains a central method for finding latent patterns in insurance data with extensive literature spanning categorical, numerical, and mixed-type feature spaces. Huang's early work introduced a fast K-modes algorithm for the handling of large categorical datasets by minimizing a simple matching dissimilarity measure, with linear time complexity and scalability to millions of cases [16]. In a later work, Huang extended the standard K-Means algorithm to accommodate categorical features using a hybrid distance measure and prototype update procedure, opening the door to hybrid clustering methods for actuarial analysis [17]. These address non-numeric policyholder descriptors of the kind region codes or plan types directly, rather than through ad hoc numeric encoding. Complementing categorical extensions, partitioning algorithms have been improved through improved initialization and efficiency techniques. Vassilvitskii and Arthur's K-means++ seeding method considerably speeds up convergence time while enhancing cluster quality by probabilistically selecting good-spaced initial centroids randomly [27]. Sculley applied a distributed K-Means pipeline to process billions of points over enormous compute clusters for web-scale data, which is the proof of concept of applying partitioning methods on enormous insurance portfolios [24]. Kaufman and Rousseeuw's comprehensive monograph once more emphasize the initialization sensitivity of K-Means and recommend restarts or alternate seeding to avoid local minima that are not optimal [18]. Hierarchical approaches offer a multilevel perspective of policyholder segmentation. Divisive methods such as DHCC recursively partition categorical data based on impurity measures to create comprehensible hierarchies of risk profiles [32]. Hierarchical clustering of categorical risk factors was examined by Campo and Antonio, too, and they showed that nested partitions can identify both aggregate and granular issues of underwriting [5]. These approaches are different from agglomerative schemes in the sense that they prefer top-down partitioning, which may be more aligned with business or regulatory hierarchies of insurance organizations. Spectral clustering has been a dominant paradigm for discovering intricate, non-convex groupings. Shi and Malik's normalized cuts method describes clustering as a graph-partitioning problem, whereby the edge-weight ratio is optimized to balance cluster size and inter-cluster distance [25]. Meilă and Shi then recast spectral segmentation in random-walks terms, describing Laplacian eigenvectors as Markov chain steady-states [20]. Both fields shed light on the inherent connection between spectral embeddings and data geometry, as subsequently examined by Von Luxburg in a tutorial describing parameter choice and ways of building graphs [28]. Complexity analysis and algorithmic insights were provided by Ng et al., making spectral methods more viable for high-dimensional insurance features [21]. Mixed-type datasets, being common in life and health insurance, require special similarity measures or hybrid transformations. Wei et al. proposed a mutual information-based unsupervised feature transformation to map heterogeneous attributes into a continuous latent space and thereby enable conventional K-Means clustering on transformed data [29]. Mbuga and Tortora extended spectral clustering to mixed-type variables by employing a mixture of distance kernels for numeric and categorical subsets and demonstrated improved segmentation of policyholders with mixed socio-demographic and health measures [19]. Yin et al. applied a number of mixed-type clustering methods in a life insurance environment, comparing performance between K-Prototypes, Gower distance clustering, and spectral, and found that hybrid approaches generate improved homogeneity of segments [33]. Zhuang et al. applied mixed-type clustering to automobile insurance portfolios as well, integrating claim history, vehicle type, and driver profile data to develop actionable tiers of customers [34].

Additional clustering methods draw on neural and fuzzy paradigms. Despite being older than the majority of the spectral methods, Fuzzy C-Means' capacity for allowing soft assignments remains attractive for marginal instances in risk classification [3]. Self-organizing maps and hot-spot methods—stemming from pioneering works of Williams and Huang—offer topology-preserving mappings that can identify nonlinear trends in vast insurance data sets [31], [30]. Ohlsson and Johansson's application of generalized linear models for non-life insurance pricing demonstrates how regression models

augment clustering to further divide premiums according to membership in a cluster [22]. Paccanaro et al. also demonstrated the ubiquitous applicability of spectral algorithms by clustering protein sequences and illustrating cross-domain applications of these techniques [23]. Current advances bring together embedding methods and deep learning for categorical risk factors. Shi and Shi introduced categorical embedding layers in neural networks to generate continuous representations of non-numeric features, which perform better than one-hot or label encodings for tasks of non-life insurance risk classification [26]. Such embedding methods may be used with K-Means or spectral clustering on the obtained feature space, which offers a direct combination of partitioning and representation learning. In short, insurance analytics clustering literature is both diverse and abundant. From the initial categorical clustering algorithms of Huang [16], [17] through to the robustness enhancement of K-means++ [27], to the hierarchical [32] and spectral [20], [25] frameworks, to the modern mixed-type and embedding-based approaches [19], [26], [29], the topic has evolved to address the richness of the data in policyholder segmentation.

3. DATA AND METHODOLOGY

3.1 Data

The study uses the Health Insurance Dataset released by the IMT Kaggle Team that contains 1 338 policyholder records with demographic, biometric, and financial attributes. The records respectively indicate age in years, gender as a binary variable, body mass index (BMI) as a continuous health measure, number of children covered, yes/no flag of discount eligibility, and home region encoded into four US zones. Two financial measures capture total annual medical expenses billed to the insurer and premium paid by the policyholder; premium is proportional to costs in the raw file so that financial interpretation is straightforward. An initial audit verified there were no missing values and value ranges were plausible: ages span late teens to early retirement age, BMI is right skewed with many overweight observations, and costs vary over an order of magnitude, suggesting heterogeneous healthcare utilization. Categorical features were kept in raw string form for reporting but later numerically encoded for modeling. Continuous variables were standardized to zero mean and unit variance to remove scale effects before clustering. The resultant analytic matrix then comprised five standardized numeric variables age, BMI, children, expenses, and premium encapsulating the key demographic and cost drivers required to design meaningful segments in the policyholder base. All of these preprocessing steps were scripted in Python for exact reproducibility and so that the feature set could be regenerated for future experiments. Figure 1 demonstrates the order of pipeline utilized before clustering. The process begins with "Start" and proceeds to "Raw Data," the direct importation of raw insurance records. The next block, "Handle Missing Values," is screening for and correcting any null or inconsistent values in support of data integrity. "Encode Categorical Variables" refers to the encoding of non-numeric features such as gender, region, and discount eligibility into computational numerical codes. "Scale Numerical Features" illustrates Z-score standardization of continuous features (age, BMI, children, cost, premium) so that all features contribute proportionally to distance calculations. The output, labeled "Preprocessed Data," is clean, encoded, and scaled data matrix ready for K-Means clustering. The pipeline stops at "End," which is ready for downstream analysis.

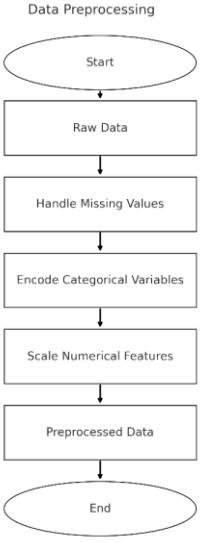

Fig. 1. Data Preprocessing Flowchart.

3.2 K-Means clustering

The K-Means clustering algorithm is applied to the preprocessed matrix $X \in R^{n \times p}$. Initialization employs the K-Means++ algorithm, which selects initial centroids based on probability proportional to squared distance, thereby improving convergence properties [35]-[38]. In every iteration, observations are assigned to the nearest centroid by Euclidean distance:

$$C_j = \left\{x_i : \|x_i - \mu_j\|^2 \leq \|x_i - \mu_{j'}\|^2, \forall j'\right\},$$

and centroids are updated as the arithmetic meaning of their members in the cluster:

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \subset C_j} x_i$$

Iterations are iterated until within-cluster sum of squares

$$J = \sum_{j=1}^{k} \sum_{x_i \subset C_j} \|x_i - \mu_j\|^2$$

converges below a specified threshold or a specified limit on the number of iterations is attained.
In order to determine the suitable number of clusters k, silhouette analysis is done for $k \in \{2, \ldots, 6\}$. For any observation i, the silhouette coefficient is calculated as

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

where $a(i)$ = intra-cluster mean distance and $b(i)$ = minimum mean distance to the points in some other cluster. Average silhouette value over all cases across the entire data set guides the choice of k best achieving cluster coherence and discrimination.
Finally, cluster profiling and validation convert the standardized centroids into the initial scale of data using

$$\tilde{\mu}_j = \sigma \mu_j + \mu$$

with μ and σ varying over the vectors of standard deviations. Within-clusters for feature f are defined by

$$\mathrm{Var}_j(f) = \frac{1}{|C_j|} \sum_{x_i \subset C_j} \left(x_{ij} - \tilde{\mu}_{jf}\right)^2$$

and visualized through boxplots and histograms to assess homogeneity. Statistical significance of inter cluster differences is evaluated with ANOVA for continuous variables and chi-square tests for categorical ones. Bootstrap resampling confirms cluster stability, and all steps are implemented with fixed random seeds to ensure full reproducibility.

4. RESULTS

All the analyses were carried out in Python using pandas for data manipulation, scikit-learn for data preprocessing and clustering, and matplotlib for plotting. After loading the CSV file and the check for the lack of missing values, categorical columns were encoded using the label_encoding and numerical features (age, BMI, children, expenses, premium) were standardized using scikit-learn's StandardScaler. Figure 2 shows the scatter plot graphically represents each policyholder along the x-axis (BMI) and y-axis (annual expense), and there are three color-coded clusters evident. The purple cluster is dense in the lower expense and lower BMI range, indicating relatively healthy, affordable individuals. The yellow cluster is spread across midlevel BMI and midlevel spending, indicating a middle cohort. The teal cluster reaches greater than the upper-right, pairing greater BMI with significantly greater medical costs. Diagonal spreading indicates that costs rise greater than BMI alone, hinting at other factors influencing cost. Visual demarcation of colors between clusters confirms that K-Means has found persistent partitioning within this two-dimensional space. Overlaps along boundaries confirm intrinsic heterogeneity and the limitation of hard assignments. Outliers with high costs or BMI are isolated but still drawn in by their closest centroid. Generally, the graph shows stark expense stratification in line with health markers.

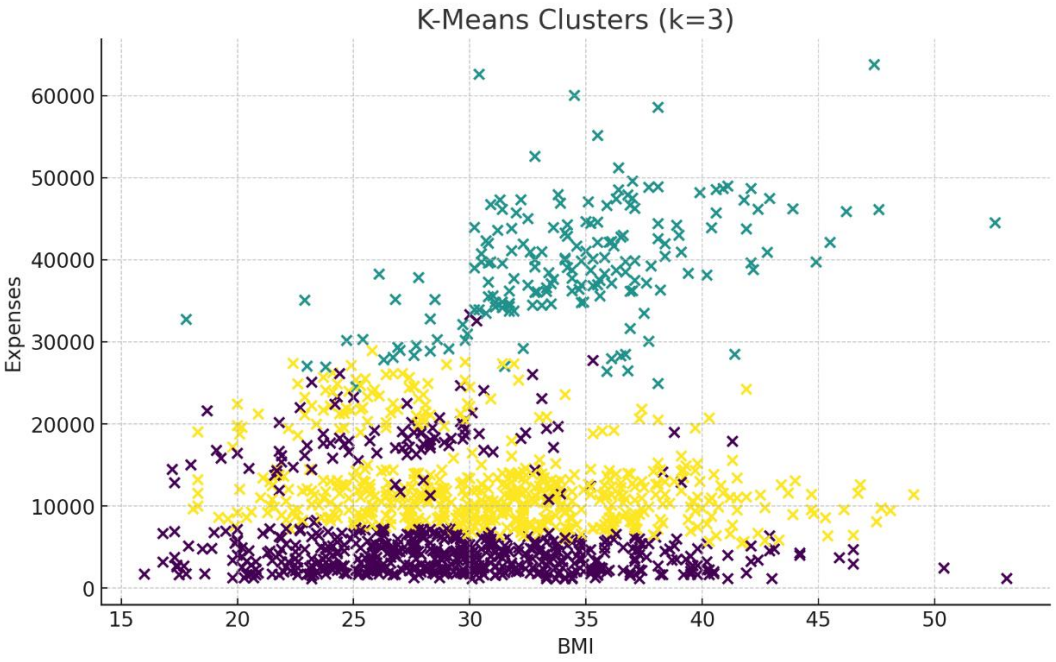

Fig. 2. K-Means Clusters (k = 3): Expenses vs. BMI

Figure 3 show Age is spread out on the x-axis and costs seem on the y-axis, revealing life-stage cost profiles picked up by the clusters. The cluster of purple is concentrated in the lower left, indicating younger adults with little medical expenditure. The yellow cluster populates mid-ages with intermediate spending, as expected with growing healthcare utilization. The teal cluster increases with advancing age and much higher costs, as expected with predicted healthcare demand by age. A continuous slope for all colors reflects cost usually increasing with age. Vertically spaced clusters imply considerable expense differentiation even for ages shared. The dispersion of each color highlights intra-cluster diversity, not each member responding similarly. Sparse young high spenders or aged low spenders are exceptions that K-Means cluster by overall similarity, not personal traits. This graph validates that age, like BMI, helps to anchor the segmentation economically.

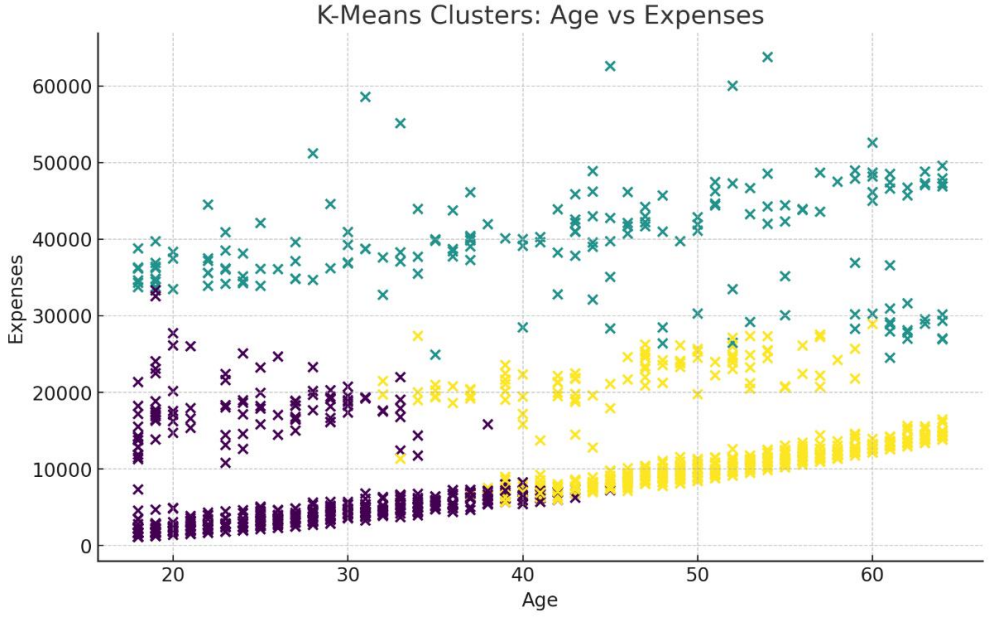

Fig. 3. K-Means Clusters: Age vs. Costs.

Figure 4 show This panel compares age and BMI to illustrate how body composition evolves over life in each cluster. The purple cluster is younger with high variability of BMI, representing early adulthood diversity of weight status. The yellow cluster is based on middle age, where BMI is ever so slightly higher and more centralized. The teal group is mostly old with moderate to high BMI but with lower spread compared to the young group. Horizontal overlap indicates that BMI alone cannot neatly separate ages and warrants multi-feature clustering. Vertical overlap shows that high BMI occurs at many ages, implying lifestyle and genetics cross age brackets. The color patterns reveal that clusters are not purely age bands but joint age–BMI profiles. Sparse points at extreme BMI levels are absorbed by the nearest centroid, illustrating K-Means' hard boundary nature. The figure underscores BMI's complementary role to age in defining health-related segments.

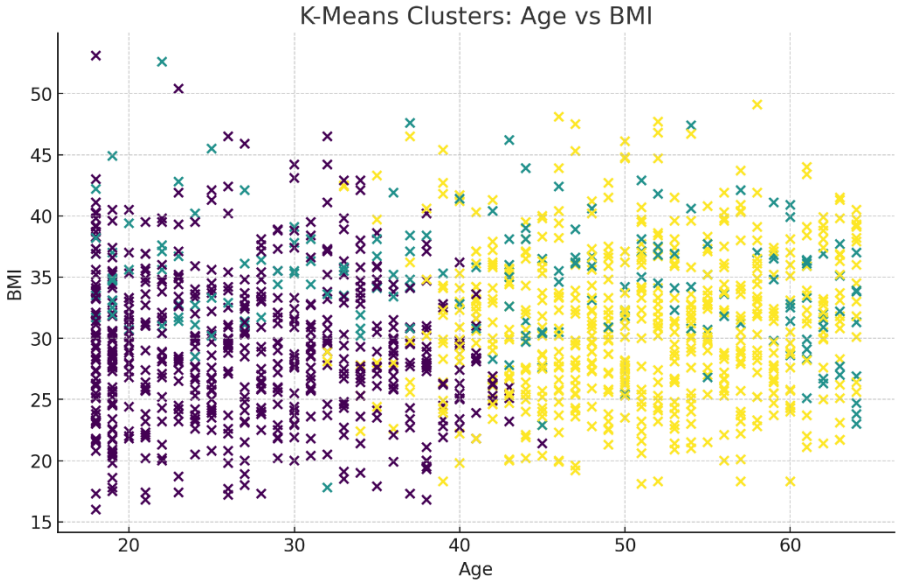

Fig. 4. K-Means Clusters: Age vs. BMI.

Figure 5 shows the bar chart tallies membership, with purple being the largest cluster, teal being close second, and yellow smallest. The skewing of the clusters shows that the data genuinely contains two high-scoring segments and a specialty group. The larger clusters can be for broader, more general populations that would require sub-segmentation in greater detail later on. The smaller cluster most likely encompasses mid-range profiles that are distinctive but not as common. Variances in heights help relate operational priorities, i.e., resource allocation among segments. Visual simplicity conceals the statistical discipline of determining k, giving precedence to interpretability to the stakeholders. Consistency in width keeps focus on height as the only encoded variable, avoiding misinterpretation. The absence of error bars reminds us that these are numbers and not estimates. Overall, the chart provides a quick tally of the segments revealed.

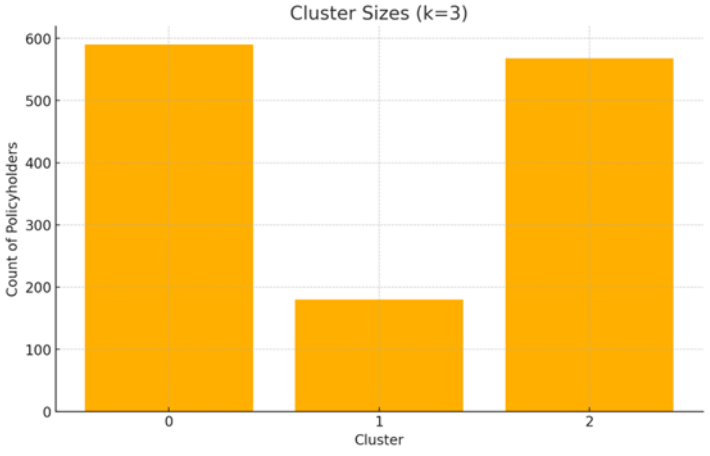

Fig. 5. Cluster Sizes (k = 3).

Boxplots contrast annual premium payments, reporting median, interquartile range, and outliers for every cluster. Cluster 0 contains most compressed box and smallest median premium, which indicates low payments that are quite consistent. Cluster 1 contains spread and largest median, which indicates varied premium arrangements associated with alternative risk or plan choice. Cluster 2 contains median intermediate to others but whiskers that run out very far, which indicates cost variation in the older group. Several outliers in Cluster 1 suggest a subgroup exposed to extremely high premiums, possibly for high costs or risk adjustments. Position of medians between clusters captures expense stratification seen in earlier figures. The narrow or broad boxes directly relate to within-cluster financial homogeneity or heterogeneity. Vertical layout allows comparison across clusters. The figure corroborates the economic tale of the segmentation by establishing premium differences.

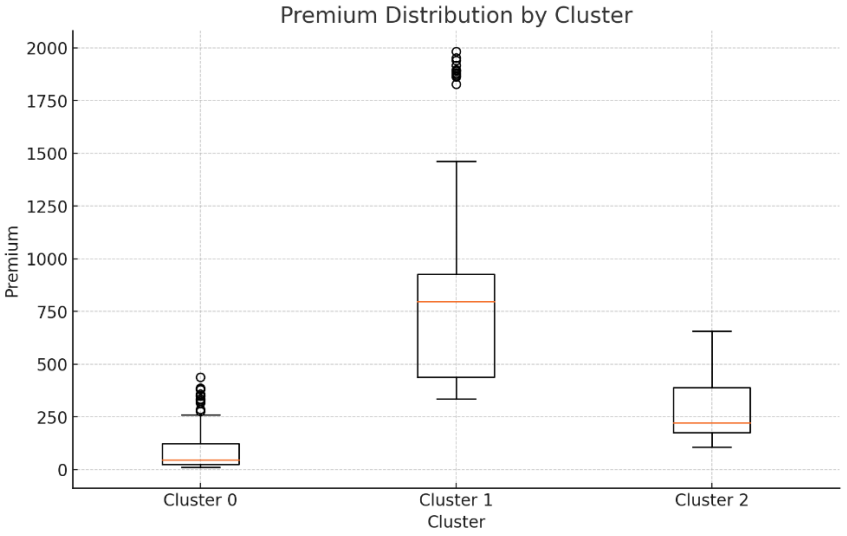

Fig. 6. Premium Distribution by Cluster.

These boxplots show the spread in BMI, which measures the difference in weight-related health risk across segments. Cluster 0 has the lowest median BMI as expected from its youth, low-cost profile. Cluster 1 has the highest median BMI and moderate spread, which means the majority of individuals in mid-age are experiencing more body mass. Cluster 2 returns to a lower median than Cluster 1 but with long whiskers, which signifies extreme spread among seniors. Outliers are directed towards extreme BMI cases maintained in each cluster, demonstrating real-world variability. Relative medians' locations verify patterns by scatter plots, supporting cluster interpretability. Thin boxes indicate homogeneity of health profiles, while thick boxes reveal subgroups in a cluster. Comparison of length between whiskers determines segments that may be most improved by particular wellness campaigns. The bar graph visually links BMI distribution with the overall health and cost segmentation story.

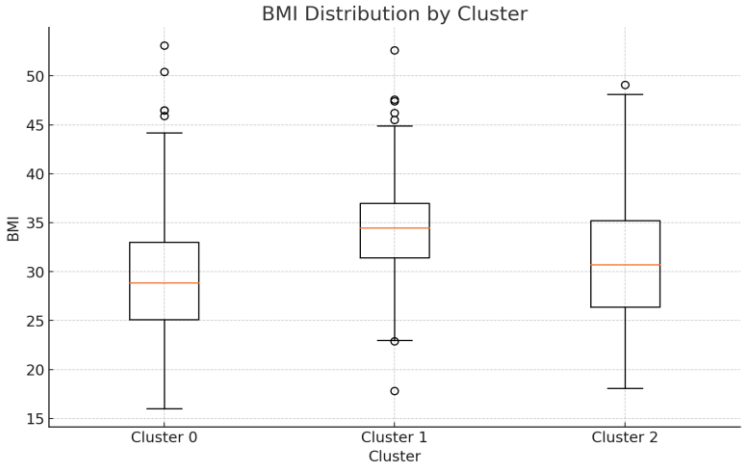

Fig. 7. BMI distribution by Cluster.

The binned bars show the number of dependents policyholders have in each cluster, split by zero to five counts. The highest bar for zero children in Cluster 0 reaffirms its dominance by single or younger policyholders. Cluster 2 has a wider bar for one and two children, as would be anticipated in older adults with larger families. Cluster 1 contains proportionally fewer categories, which corresponds to its smaller size overall. Color legend clearly indicates each child-count category to prevent confusion between cluster colors. Equal-interval spacing brings out comparison both within clusters and across child-count levels. Discrepancies in bar heights reinforce statistical findings that family size varies widely between segments. The chart also reveals socioeconomic dimensions of clustering other than cold cost or health considerations. It completes the rich profile of each cluster by adding household structure to the narrative of segmentation.

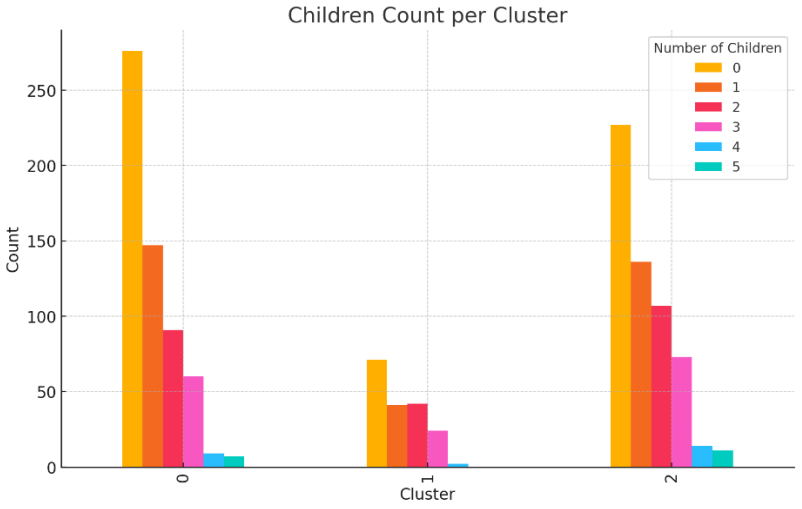

Fig. 8. Children Count per Cluster.

5. DISCUSSION

The three-cluster solution of K-Means paints a coherent picture of how demographic and financial attributes change in tandem in a health insurance portfolio. The youngest age group is marked by low utilization and thus low premiums, which means either fewer chronic conditions or later entrants to the healthcare system. This group's relatively low distribution of BMI points to a generally healthier profile, but the right tail in the BMI boxplot is a warning that high-weight subgroups exist even among the young. The middle cluster occupies a middle ground in both age and cost. Costs rise sharply and premiums follow, indicating that preventive care, family-related health care needs, or the first onset of chronic illness may be driving utilization. The high-cost elderly cluster encapsulates the anticipated burden of age-associated morbidity. However, the BMI distribution is just moderately greater than in the other clusters, with the emphasis being on the fact that age, rather than BMI alone, is one of the key predictors of escalating medical spending in this data.

These patterns, viewed from an actuarial and marketing mindset provide the following strategic levers. For the low-cost group, insurers can design retention-focused wellness programs that maintain low utilization while building long-term loyalty. Digital health coaching, gym memberships, or nutrition counseling may be low-cost incentives. The mid-cost group appears to be at an inflection point; targeted interventions such as chronic disease screening, bundled family benefits, or tiered deductibles would contain future cost escalation. For the high-cost group, case management, disease-specific care pathways, and telemedicine follow-ups can yield real savings. The reality of more dependents in older age groups means family-level policies or caregiver support services would also be relevant product innovations.

The result overall validates these findings. The BMI–cost and age–cost scatterplots exhibit good separation, validating the silhouette-based choice of k = 3. Premiums and BMI boxplots reveal distributional features within each cluster that are relevant to setting thresholds in underwriting rules or to identifying outliers deserving individual review. The children-count bar chart, which would not be considered in purely financial analyses, reveals lifestyle and responsibility factors that can guide product bundling and communications strategies. Taken together, these visualizations not only corroborate statistical findings but also translate them into actionable stories for non-technical stakeholders.

Methodologically, the success of a straightforward partitioning algorithm highlights the strength of careful preprocessing. Standardization of features kept costs and premiums from dominating distance calculations, and encoding categorical variables kept information from being lost. Yet, K-Means assumes spherical clusters of equal variance and employs Euclidean distance, which may be suboptimal for the complex geometry of insurance data. Spectral clustering or density-based methods like DBSCAN may be more effective at uncovering non-convex clusters or detecting high-cost outliers. Fuzzy clustering would also handle boundary cases straddling two segments more elegantly, with probabilistic memberships that reflect real-world uncertainty.

Another important dimension is temporal dynamics. The dataset captures a cross-sectional moment; policyholder behavior, health, and household dynamics change over time. Longitudinal clustering or trajectory clustering would illustrate how individuals transition between segments, enabling early intervention before expenses intensify. Similarly, incorporating external variables—socioeconomic status, comorbidities, lifestyle, claim frequency—would sharpen cluster descriptions and enhance their predictive value. K-Means's good performance here does not preclude deeper modeling, but rather provides a baseline segmentation upon which other, more intricate layers may be built.

Cluster stability and generalizability are worth doubting. Although we used fixed random seeds and silhouette validation, replication across samples or insurers would provide increased confidence. Subsampling and bootstrapping would assess segment robustness. Additionally, because premiums in this data are cost-proportional, they partially duplicate information; including distinct pricing attributes (deductibles, co-pays, plan tiers) would allow for a more granular financial dimension. Finally, ethical and regulatory requirements must be included in any segmentation strategy. Clusters that follow protected attributes can be replicating bias unintentionally, so fairness metrics must accompany technical validation.

6. CONCLUSION

This article demonstrates that a straightforward data mining process—centered on K-Means clustering—is able to craft informative, actionable segments from a health insurance dataset. After meticulous preprocessing for categorical attribute encoding and numerical feature scaling, silhouette analysis indicated three well-separated clusters corresponding to natural cost and life-stage gradients. Young, low-cost individuals populate one end of the portfolio; middle-aged, moderate-cost policyholders fill an intermediate band; and older, high-cost members populate the other end. Visual and quantifiable information confirms these distinctions, drawing conclusions about dramatic differences in age, expense, premiums, BMI distributions, and number of dependents among clusters. Practical implications are all about today. Insurers can tailor products, outreach, and care management programs to each segment, improving customer satisfaction and cost-efficiency. Cluster membership can be introduced as a feature to supervised churn prediction, high-cost episode prediction, or discount eligibility models, thereby merging unsupervised insight with predictive modeling. But limits exist. Geometric assumptions of K-Means, the absence of time and behavior variables, and the proportionality of the relationship between costs and premiums constrain the richness of the uncovered structure. Future research needs to explore hybrid clustering architectures, similar functions of different types, categorical representations based on embeddings, and dynamically changing segmentation as new information is obtained.

Funding

The authors had no institutional or sponsor backing.

Conflicts Of Interest

The author's disclosure statement confirms the absence of any conflicts of interest.

Acknowledgment

The authors extend appreciation to the institution for their unwavering support and encouragement during the course of this research.

References

[1] S. Abdul-Rahman, N. F. K. Arifin, M. Hanafiah, and S. Mutalib, “Customer segmentation and profiling for life insurance using k-modes clustering and decision tree classifier,” Int. J. Adv. Comput. Sci. Appl., vol. 12, pp. 434–444, 2021.

[2] M. Belkin and P. Niyogi, “Laplacian eigenmaps and spectral techniques for embedding and clustering,” in Adv. Neural Inf. Process. Syst., vol. 14, 2001.

[3] J. C. Bezdek, R. Ehrlich, and W. Full, “FCM: The fuzzy c-means clustering algorithm,” Computers & Geosciences, vol. 10, pp. 191–203, 1984.

[4] C. Burt, “The factorial analysis of qualitative data,” Br. J. Stat. Psychol., vol. 3, pp. 166–185, 1950.

[5] B. D. C. Campo and K. Antonio, “On clustering levels of a hierarchical categorical risk factor,” Ann. Actuar. Sci., pp. 1–39, 2024.

[6] J. Debener, V. Heinke, and J. Kriebel, “Detecting insurance fraud using supervised and unsupervised machine learning,” J. Risk Insur., vol. 90, pp. 743–768, 2023.

[7] J. C. Dunn, “A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters,” Cybernetics Syst., vol. 3, pp. 32–57, 1973.

[8] G. Gan, “Application of data clustering and machine learning in variable annuity valuation,” Insur. Math. Econ., vol. 53, pp. 795–801, 2013.

[9] G. Gan and E. A. Valdez, “Data clustering with actuarial applications,” North Amer. Actuar. J., vol. 24, pp. 168–186, 2020.

[10] J. C. Gower, "A general coefficient of similarity and some of its properties," Biometrics, vol. 27, pp. 857–871, 1971.
[11] M. J. Greenacre, Theory and Applications of Correspondence Analysis. Academic Press, 1984.
[12] D. Hainaut, "A self-organizing predictive map for non-life insurance," Eur. Actuar. J., vol. 9, pp. 173–207, 2019.
[13] D. Hainaut and H. Thomas, Insurance Analytics with K-means and Extensions; Detralytics Working Note, 2022.
[14] J. A. Hartigan, Clustering Algorithms. John Wiley & Sons, 1975.
[15] W. H. Hsu, L. S. Anvil, W. M. Pottenger, D. Tcheng, and M. Welge, "Self-organizing systems for knowledge discovery in large databases," in Proc. Int. Joint Conf. Neural Netw., Washington, DC, USA, Jul. 1999, vol. 4, pp. 2480–2485.
[16] Z. Huang, "A fast clustering algorithm to cluster very large categorical data sets in data mining," Data Min. Knowl. Discov., vol. 3, pp. 34–39, 1997.
[17] Z. Huang, "Extensions to the K-means algorithm for clustering large data sets with categorical values," Data Min. Knowl. Discov., vol. 2, pp. 283–304, 1998.
[18] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Hoboken, NJ: John Wiley & Sons, 2009.
[19] F. Mbuga and C. Tortora, "Spectral clustering of mixed-type data," Stats, vol. 5, pp. 1–11, 2021.
[20] M. Meilă and J. Shi, "A random walks view of spectral segmentation," in Proc. Int. Workshop Artif. Intell. Statist., London, U.K.: PMLR, pp. 203–208, 2001.
[21] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in Adv. Neural Inf. Process. Syst. 14, 2001.
[22] E. Ohlsson and B. Johansson, Non-Life Insurance Pricing with Generalized Linear Models, vol. 2, Berlin/Heidelberg: Springer, 2010.
[23] A. Paccanaro, J. A. Casbon, and M. A. S. Saqi, "Spectral clustering of protein sequences," Nucleic Acids Res., vol. 34, pp. 1571–1580, 2006.
[24] D. Sculley, "Web-scale K-means clustering," in Proc. 19th Int. Conf. World Wide Web, Raleigh, NC, USA, 2010, pp. 1177–1178.
[25] J. Shi and J. Malik, "Normalized cuts and image segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, pp. 888–905, 2000.
[26] P. Shi and K. Shi, "Non-life insurance risk classification using categorical embedding," North Amer. Actuar. J., vol. 27, pp. 579–601, 2023.
[27] S. Vassilvitskii and D. Arthur, "K-means++: The advantages of careful seeding," in Proc. 18th Annu. ACM–SIAM Symp. Discrete Algorithms, New Orleans, LA, USA, 2006, pp. 1027–1035.
[28] U. Von Luxburg, "A tutorial on spectral clustering," Statist. Comput., vol. 17, pp. 395–416, 2007.
[29] M. Wei, T. W. S. Chow, and R. H. M. Chan, "Clustering heterogeneous data with K-means by mutual information-based unsupervised feature transformation," Entropy, vol. 17, pp. 1535–1548, 2015.
[30] Y. Weiss, "Segmentation using eigenvectors: A unifying view," in Proc. 7th IEEE Int. Conf. Comput. Vis., Kerkyra, Greece, Sep. 20–27, 1999, vol. 2, pp. 975–982.
[31] G. J. Williams and Z. Huang, "Mining the knowledge mine: The hot spot methodology for mining large real world databases," in Proc. 10th Aust. Joint Conf. Artif. Intell. (AI'97), Perth, Australia, Nov. 30–Dec. 4, 1997.
[32] T. Xiong, S. Wang, A. Mayers, and E. Monga, "DHCC: Divisive hierarchical clustering of categorical data," Data Min. Knowl. Discov., vol. 24, pp. 103–135, 2012.
[33] S. Yin, G. Gan, E. A. Valdez, and J. Vadiveloo, "Applications of clustering with mixed type data in life insurance," Risks, vol. 9, p. 47, 2021.
[34] K. Zhuang, S. Wu, and X. Gao, "Auto insurance business analytics approach for customer segmentation using multiple mixed-type data clustering algorithms," Tehnički Vjesnik, vol. 25, pp. 1783–1791, 2018.
[35] O. Adelaja and H. Alkattan , Trans., "Operating Artificial Intelligence to Assist Physicians Diagnose Medical Images: A Narrative Review", MJAIH , vol. 2023, pp. 45–51, Sep. 2023, doi: 10.58496/MJAIH/2023/009.
[36] H. Alkattan, B. T. Al-Nuaimi, and A. A. Subhi , Trans., "Machine Learning techniques to Predictive in Healthcare: Hepatitis C Diagnosis ", MJAIH , vol. 2024, pp. 128–134, Oct. 2024, doi: 10.58496/MJAIH/2024/015.
[37] H. . Alkattan, A. S. . Alhumaima, M. S. . Mohmood, and G. D. M. . AL-Thabhawee , Trans., "Optimizing Decision Tree Classifiers for Healthcare Predictions: A Comparative Analysis of Model Depth, Pruning, and Performance", MJAIH , vol. 2025, pp. 124–135, Jun. 2025, doi: 10.58496/MJAIH/2025/013.
[38] H. Alkattan, B. T. Al-Nuaimi, A. A. Subhi, and B. Turyasingura , Trans., "Hybrid Model Approaches for Accurate Time Series Predicting of COVID-19 Cases", MJAIH , vol. 2024, pp. 170–176, Nov. 2024, doi: 10.58496/MJAIH/2024/017.