

## Health Insurance Risk Segmentation Using Unsupervised Machine Learning and Clustering Techniques

تصنيف مخاطر التأمين الصحي باستخدام تقنيات التعلم الآلي والتجميع غير الخاضعة للرقابة

Taqwa Salim Mohammed تقوى سالم محمد

كلية الطب، جامعة ديالى، ديالى، العراق

taqwa@uodiyala.edu.iq

تاريخ التعديل : 2026/06/29

تاريخ النشر : 2026/10/25

تاريخ تقديم البحث : 2026/06/07

تاريخ قبول النشر : 2026/08/23

### Abstract

Health insurance databases include key demographic, health-related, lifestyle and socioeconomic risk factors that allow for the discovery of patterns of policyholder risk. This paper applies unsupervised machine learning techniques to analyze an insurance cost and risk dataset of 1,338 rows/12 variables. In particular, the overall goal is to find hidden risk classes among policyholders and gain better insights into determinants of differences in insurance premiums. Initial preprocessing included checking variable types, imputing missing numerical values, encoding categorical variables, and standardizing numerical features. Next, multiple clustering techniques were used which are BIRCH, Gaussian Mixture Model, Agglomerative Clustering (average linkage), Agglomerative Clustering (complete linkage) and DBSCAN. Clusters were evaluated using internal validation metrics, namely Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score. Cluster sizes were examined, and the results were visualized using two-dimensional PCA plots. Results: BIRCH provided the best clustering framework that clustered all data into three identifiable low, medium and high-risk insurance groups. The main differences between insurance risk groups were seen in age, BMI, insurance premiums and associated health-risk factors based on cluster profiling. Compared to the low-risk group, policyholders in the high-risk group had higher average insurance charges, older age, and a greater burden of obesity. The findings nonetheless suggest that techniques for clustering are able to partition health insurance data, thereby enabling decisions based on modelling and evidence on rate-setting, price elasticity in the insurance context, and rationalizing expenditures in a healthcare system.

**Key words:** Health insurance, clustering, BIRCH, Gaussian Mixture Model, DBSCAN, risk segmentation, data preprocessing, unsupervised machine learning.

### المستخلص

تتضمن قواعد بيانات التأمين الصحي مجموعة من عوامل الخطر الرئيسية، تشمل العوامل الديموغرافية والصحية ونمط الحياة والعوامل الاجتماعية والاقتصادية، مما يسمح باكتشاف الأنماط المرتبطة بالمخاطر لدى حاملي وثائق التأمين. تستخدم هذه الدراسة تقنيات التعلم الآلي غير الخاضع للإشراف لتحليل مجموعة بيانات خاصة بتكاليف التأمين ومخاطره، تتكون من 1338 سجلاً و12 متغيراً. ويتمثل الهدف العام، على وجه الخصوص، في الكشف عن فئات المخاطر الكامنة بين حاملي وثائق التأمين، والتوصل إلى فهم أعمق للعوامل المحددة للاختلافات في أقساط التأمين. شملت مرحلة المعالجة المسبقة الأولية التحقق من أنواع المتغيرات، وتعويض القيم العددية المفقودة، وترميز المتغيرات الفئوية، وتوحيد قياس الخصائص العددية. بعد ذلك، استُخدمت عدة تقنيات للتجميع العنقودي، هي: خوارزمية BIRCH، ونموذج المزيج الغاوسي (Gaussian Mixture Model)، والتجميع الهرمي التكتلي باستخدام الربط المتوسط (Average Linkage)، والتجميع الهرمي التكتلي باستخدام الربط الكامل (Complete Linkage)، وخوارزمية DBSCAN. وجرى تقييم العناقيد باستخدام مقاييس التحقق الداخلي، وهي: معامل السيلويت (Silhouette Score)، ومؤشر ديفيز-بولدين (Davies-Bouldin Index)، ومؤشر كالينسكي-هاراباز (Calinski-Harabasz Score).

(Harabasz Score). كما جرى فحص أحجام العناقيد، وتمثيل النتائج بصرياً باستخدام مخططات ثنائية الأبعاد لتحليل المكونات الرئيسية (PCA). النتائج: أظهرت خوارزمية BIRCH أفضل إطار عمل للتجميع العنقودي، إذ صنفت البيانات بأكملها إلى ثلاث مجموعات تأمينية واضحة ومتميزة وفق مستوى المخاطر: منخفضة ومتوسطة ومرتفعة. واستناداً إلى توصيف العناقيد، تمثلت أبرز الاختلافات بين مجموعات المخاطر التأمينية في العمر، ومؤشر كتلة الجسم (BMI)، وأقساط التأمين، وعوامل المخاطر الصحية المرتبطة بها. وبالمقارنة مع مجموعة المخاطر المنخفضة، اتسم حاملو وثائق التأمين ضمن مجموعة المخاطر المرتفعة بارتفاع متوسط تكاليف التأمين، وتقدم العمر، وارتفاع معدلات السمنة. وتشير النتائج، مع ذلك، إلى أن تقنيات التجميع العنقودي قادرة على تقسيم بيانات التأمين الصحي إلى مجموعات ذات خصائص متميزة، بما يتيح دعم القرارات المستندة إلى النمذجة والأدلة فيما يتعلق بتحديد معدلات التأمين، ومرونة الأسعار في السياق التأميني، وترشيد النفقات في نظام الرعاية الصحية.

**الكلمات المفتاحية:** التأمين الصحي، التجميع العنقودي، خوارزمية BIRCH، نموذج المزيج الغاوسي، خوارزمية DBSCAN، تجزئة المخاطر، المعالجة المسبقة للبيانات، التعلم الآلي غير الخاضع للإشراف.

## 1. Introduction

Health insurance has emerged as a key means of paying for health care in the modern world because it offers financial protection against unexpected medical events and enables access to care. At the same time, insurance companies face a perennial challenge to understand how different demographic, lifestyle, medical and socioeconomic factors drive the costs and risks of insurance. Age, body mass index, smoking status, blood pressure, prior diseases and disorders of the lungs and heart or just about any background condition that might affect their health (along with occupational characteristics), exercise intensity or frequency, income level all can combine in various ways to characterize a policyholder. As a result, health insurance analysis has moved beyond simple descriptive statistics and cost summaries. State-of-the-art data-driven approaches are increasingly and systematically used to reveal hidden patterns in the data, target at-risk populations at high risk of developing chronic diseases, and thus enabling a much-needed transparency of decision-making processes in health care or insurance systems.

More recent studies show that machine learning is an effective means for health insurers to price and analyze relevant costs by highlighting features affecting medical expenditures as well as segments of consumers. Recent work on intelligent health insurance pricing systems has used both machine learning and risk-based frameworks to improve premium estimates and decision support [1]. The research on personalization of insurance premiums has brought out the necessity of machine learning and explainable analytics in understanding consumer segmentation and influencing factors for pricing the insurance [2]. These studies show that health insurance prices are determined by a variety of factors, including health, behavioral, demographic and socioeconomic characteristics. It requires systematic analytical criteria for segmenting policyholders into major segments and characterizing the features of each segment.

The majority of health insurance studies discussed above are focused on supervised learning, where the primary goal is to predict a target variable in this case insurance premiums or claim expenses. Even though prediction under supervision is useful, it might not reveal the complete internal structure of the dataset if class labels are not provided. In various real-world insurance applications, the goals are not simply to predict cost, but rather to identify how policyholders naturally group together based on their shared characteristics. This makes unsupervised learning highly beneficial. Unsupervised learning is designed to work with unlabeled data, and it is used to find hidden structures, patterns or clusters without relying on predefined class labels [3]. This approach has been used frequently in healthcare for patient segmentation, risk stratification and phenotyping, particularly when clinical or administrative data include heterogeneous populations [4]. Clustering is one of the most common approaches in unsupervised learning. The main goal is to form groups whose members are more similar to one another than to members of other groups. Clustering has been used in healthcare research to establish the composition of patient cohorts,

identify hidden clinical patterns, and augment personalized decision support. The use of clustering methods to describe biologically important subgroups and the potential for them to help understand complicated health data was highlighted in a recent paper using unsupervised learning methods for precision medicine [5]. A practical application of clustering in insurance is the segmentation of policyholders into low, mid and high risk based on similarities found in demographic, health, behavioral or financial data. This segmentation can help insurance analysts identify the factor interaction associated with high insurance costs and risk profiles that are potentially low predicted risk. This study applies clustering algorithms to a health insurance dataset containing 1,338 records and 12 variables. The dataset contains demographics (age, sex, region and number of children), health-related variables (BMI, blood pressure and pre-existing conditions), lifestyle-related variables (exercise frequency) and occupational risk data; one socioeconomic variable defined by annual income; and finally, the related insurance costs set by charges. The dataset was systematically pretreated before clustering. This involved checking data types, identifying missing values, separating numerical and categorical variables, imputing missing numerical values, encoding categorical features, and standardizing numerical features. Clustering algorithms hinge on these procedures as they are sensitive to feature scaling, completeness of data and representing categorical variables. This paper evaluates five clustering approaches (BIRCH, Gaussian Mixture Model, Agglomerative Clustering with average linkage, Agglomerative Clustering with complete linkage, DBSCAN). These methods were selected as they serve to demonstrate some of the different clustering techniques. BIRCH is an in-memory, hierarchical-clustering algorithm suitable for incremental or large-scale clustering tasks that produces a tree structure to represent the inherent data by forming hierarchies [6]. Gaussian Mixture Model (GMM) is a probabilistic clustering algorithm that models the data as a mixture of several Gaussian distributions and can be useful in cases where there may be overlapping clusters. Agglomerative clustering is performed with a bottom-up hierarchical approach combining data at each iteration based on linkage distance [8]. DBSCAN is a density-based algorithm that can identify dense regions and noise points without requiring the number of clusters in advance. In this study, we perform multiple methodologies over the same health insurance dataset to analyze tree, probabilistic, hierarchical and density-based behavior of grouping. In unsupervised analysis, where there is no initial target label for verifying that a cluster is accurate without ambiguity, simple metrics won't suffice; it needs some form of cluster validation. As a result, the paper evaluates the quality of clustering results through internal validation metrics like Silhouette Score, Davies–Bouldin Index and Calinski-Harabasz Score. These measures are often used in case of lack of ground-truth labels to validate cluster compactness and separation [10]. Additionally, PCA is used to visualize clusters in 2D space. PCA visualization does not replace numerical validation, but it helps show whether the reduced clusters are separated or overlapping. The main contribution of this work is to develop a holistic framework for clustering-based health insurance risk segmentation. This work focuses on discovering naturally existing risk clusters in the dataset, rather than just treating insurance pricing as a regression task. The proposed workflow includes: pre-processing, clustering, internal validation, PCA visualization and cluster profiling followed by low-, medium- and high-risk group interpretation. It presents a more analytical and easy-to-understand viewpoint of the trends in insurance risk. This study compares several clustering approaches to identify a practical grouping structure for the heterogeneous health insurance dataset.

## 2. Related work

Recent studies show that artificial intelligence is becoming important in insurance analytics, particularly for risk assessment, underwriting, pricing, fraud detection, and claims administration. Bhattacharya et al. [11] surveyed various applications of AI in insurance and discussed that modern-day insurance is moving towards a more data-driven paradigm but highlighted important concerns around transparency, regulation, fairness, and real-world applicability. This study is relevant because health insurance risk grouping requires both computational modeling and a

readable, interpretable procedure explaining how groups of policyholders are formed. Likewise, du Preez et al. [12] studied the application of machine learning in healthcare claims analysis and showed that complicated relationships are readily identifiable from insurance-related healthcare data if only the data is accurately collected and passed through quality control. These studies confirm that insurance analytics benefits from machine learning; however, they also suggest that reliable preprocessing and interpretation are necessary before any model output can be applied in practice.

Most health insurance studies still focus on supervised prediction. Seyam [13] applied machine learning models to predict the expenditures of high-cost healthcare consumption using data from private group health insurers and demonstrated the potential for cost control and resource allocation via risk classification. Bu et al. [14] used hospital departmental indicators to improve the early detection of health insurance overspending risk. This is an interesting study because it shows that health insurance costs are largely driven by many administrative, demographic and other factors. However, the focus is mostly on prediction that suggests the model is based on a pre-determined intended target event. In contrast, this study employs unsupervised clustering to position inherent policyholder groupings without the need for initial class identification. This makes the analysis more explorative and suitable to uncover hidden low-, medium- and high-risk profiles in the dataset.

Clustering methods have become widespread in insurance analytics because they help group policyholders with similar characteristics. Jamotton et al. [15] explored some clustering methods for insurance datasets (as well as datasets relevant to actuarial research) and pointed out that real-life insurance portfolios are often heterogeneous with both categorical and numerical features. This relates directly to the health insurance data set, including gender, smoking status, location of living (urban or rural), pre-existing diseases (yes or no), which job have they in their work life, age, body mass index (BMI) and blood pressure, yearly income and also the medical cost. Due to the sensitivity of clustering methods to feature representation, mixed-type insurance data require meticulous preparation right before performing clustering. This paper uses missing-value imputation, categorical encoding and numerical normalization as standard preprocessing steps before using BIRCH, Gaussian Mixture, Agglomerative Average, Agglomerative Complete and DBSCAN methods.

Recent healthcare research supports the use of clustering to improve risk stratification. Alshrouf [16] proposed an improvement for hierarchical clustering with BIRCH, focusing on partitioning medical data, showing that BIRCH is still highly useful in the grouping of health care datasets. Onojete et al. [17] performed illness clustering based on symptoms and showed that unsupervised learning can learn meaningful relationships between medical diseases and improve the understanding of hidden health signatures. Rawat et al. [18] used PCA-based visualization and internal validation of the cluster structure after various unsupervised learning methodologies on patient records without any prediction from these clusters. The investigations are supportive of the methods of this study as population-segmentation and health insurance data can be treated in same way, which means determined cohorts of individuals are identified on similar characteristics in terms of their health and cost.

It is important to compare multiple clustering techniques because no single method is best for every dataset. Kossakov et al. [19] compared clustering methods for healthcare data using the Silhouette Score, Davies–Bouldin Index and Calinski–Harabasz Score. Abhadiomhen et al. [20] reviewed health risk assessment with machine learning, where they mentioned the key points associated with medical data collection, preprocessing, validation and explainability of healthcare applications related to risk. Chicco et al. [21] discussed clustering quality and showed that metrics like Silhouette coefficient and Davies–Bouldin index are helpful for validating unsupervised clustering. These studies support the importance of applying multiple validation criteria in the present study rather than relying solely on visual inspection.

Clustering for healthcare decision-making is strengthened by interpretable cluster profiling. Yang et al. [22] presented a clustering-based healthcare system for assessing patient risks and

trends. Beaney et al. [23] explored clustering in the context of a variety of chronic conditions and clarify that patient groups can be formed upon disease similarity, care needs or health outcomes. Zonayed et al. [24] reviewed available machine learning and IoT applications used in real-world healthcare practice, noting that unsupervised learning enables segmentation, clustering, and anomaly detection within healthcare data. They show that clustering functions not only as a mathematical grouping tool but also as a conceptual framework for understanding heterogeneous patient or policyholder characteristics. Despite these advancements, significant gaps remain in the literature. Most health insurance research focuses on direct cost prediction rather than hidden policyholder groups. Second, the preprocessing phase is still somewhat neglected in several clustering studies even though missing values, categorical variables and heterogeneous measurement scales on features can severely influence clustering results. Third, many studies report validation scores without closely investigating how the resulting groupings align with true insurance-risk characteristics. To address these shortcomings, this study executes a holistic unsupervised workflow via data inspection (quality review), preprocessing, feature preparation, clustering comparison, PCA-based visualization of the result, internal validity and cluster profiling. The long-term goal is not only to create clusters but rather to interpret them as meaningful health insurance risk classes to enable data-driven risk assessment.

### **3. Data and Methodology**

#### **3.1. Data Description**

The Health Insurance Cost and Risk Dataset contain 1,338 records and 12 variables covering demographic, health, lifestyle, socioeconomic, and insurance-cost information [25]. The dataset was downloaded in CSV format and prepared for unsupervised machine learning. The clustering variables were age, sex, BMI, number of children, smoking status, region, blood pressure, exercise frequency, pre-existing conditions, occupational risk, and annual income. The Charges variable was excluded from cluster construction and was used only after clustering to interpret cost differences among the resulting risk groups.

Before analysis, the dataset was checked for missing values, variable types and distributions. Missing values occurred only in a few numerical variables. Because clustering methods require complete data, median imputation was used to replace missing numerical values. We selected the median imputation method because it is less sensitive to extreme values than mean imputation, especially in health and insurance datasets where variables like BMI, income and charges tend to have highly skewed distributions or extreme values.

The overall methodological approach for health insurance risk grouping is shown in Figure 1. The process begins with importing the Health Insurance Cost and Risk Dataset, followed by preprocessing steps such as missing-value treatment, categorical encoding, and feature scaling. After feature preparation and exploration, five clustering methods were used: BIRCH, Gaussian Mixture Model, Agglomerative Average, Agglomerative Complete and DBSCAN. Clustering results were additionally evaluated using Silhouette Score, Davies-Bouldin Index, Calinski–Harabasz Score and examining cluster sizes. Finally, BIRCH was selected as the final clustering model, and its clusters were interpreted as low, medium and high-risk insurance profiles.

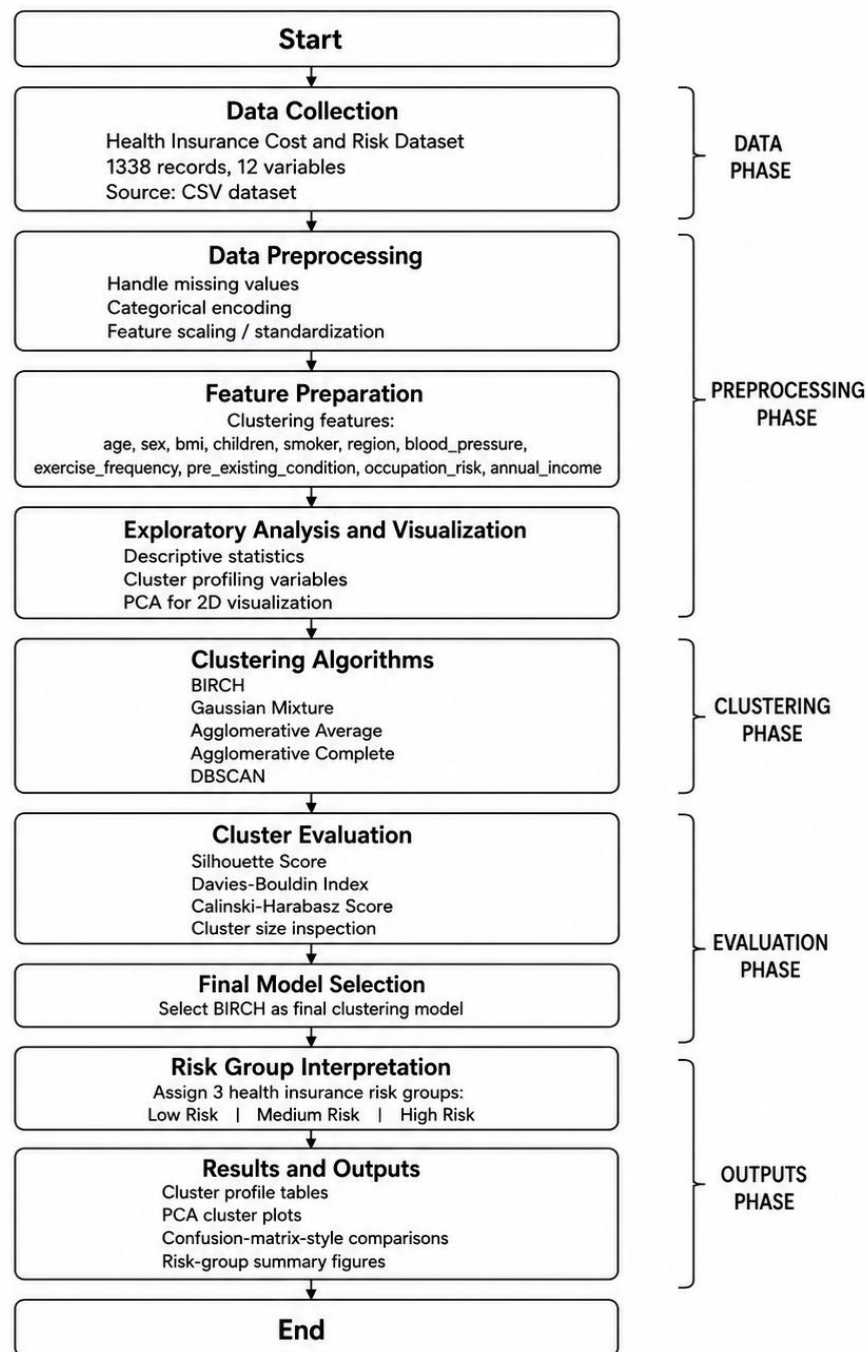


Figure 1. Flowchart of the Health Insurance Risk Clustering Framework.

### 3.2. Data Preprocessing

Data preprocessing was performed before implementing the clustering methods. This stage consisted of handling missing values, encoding categorical variables and standardizing numerical features. The preprocessing phase is important because clustering algorithms are sensitive to distance, and variables recorded on larger scales can dominate the process if they are not standardized. As an example, yearly income has much larger numerical values than age or BMI, and therefore the features were scaled before clustering so that no feature would bias the final clusters based only on their measurement scale.

For missing numerical data, median imputation was used as follows:

$$x_{ij}^* = \begin{cases} x_{ij}, & \text{if } x_{ij} \text{ is observed} \\ \text{median}(X_j), & \text{if } x_{ij} \text{ is missing} \end{cases}$$

where  $x_i$  refers to a numerical value, and  $(X)$  represents the median of that feature. This prevents missing values from affecting cluster formation.

Categorical features (e.g., sex, smoking status, region, exercise frequency, pre-existing conditions and occupational risk) were transformed into numerical format using one-hot encoding. This process converts each category to a binary indicator (1 or 0). We can summarize the encoding process as follows:

$$z_{ijm} = \begin{cases} 1, & \text{if } x_{ij} = c_m, \\ 0, & \text{otherwise} \end{cases}$$

where  $z_{ijm}$  is the encoded binary value for observation  $i$  in category  $c_m$ . This allows categorical data to be used in clustering algorithms.

After encoding, numerical features were normalized using z-score normalization. Standardization transforms each numerical feature into a distribution with zero mean and unit variance.

The numerical features were standardized using z-score normalization. Each feature therefore has a mean of 0 and a standard deviation of one:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

where  $z_{ij}$  is the standardized value,  $x_{ij}$  is the original value,  $\mu_j$  is the feature mean, and  $\sigma_j$  is the feature standard deviation. This step is important because the variables have different scales and units.

### 3.3. Methodological Framework

The methodology of this study includes five main stages, including a data collection stage, preprocessing stage, model implementation stage (clustering), validation stage and interpretation of clusters. After preprocessing, various clustering techniques were applied to the prepared dataset. The aim was to assess different unsupervised learning approaches and determine the most suitable clustering framework for segmenting health insurance risk.

The general clustering structure is stated as follows:

$$C = \{C_1, C_2, \dots, C_k\}$$

where  $C$  is the set of clusters,  $C_k$  is the  $k$ -th cluster. Policyholders with similar characteristics are assigned to the same cluster. The study used three risk groups: low, medium, and high.

### 3.4. BIRCH Clustering

BIRCH was used as the main clustering method because of its efficiency in fast and compact grouping. BIRCH uses a clustering feature tree to represent the data points as concise cluster statistics. The clustering feature can be written as follows:

$$CF = (N, LS, SS)$$

where  $N$  is the number of observations in the cluster,  $LS$  is the linear sum, and  $SS$  is the squared sum. These values are used to calculate the cluster centroid.

$$\mu = \frac{LS}{N}$$

This is an advantage of BIRCH which effectively summarizes clusters and generates compact groups. The study used BIRCH to create three clusters, which were finally classified as low-risk, medium-risk and high risk based on their similarities.

### 3.5. Gaussian Mixture Model

The Gaussian Mixture Model is a probabilistic clustering algorithm. Unlike hard clustering methods, it assumes that the data comes from a mixture of Gaussian distributions. The probability density function of a GMM is expressed as follows:

$$p(x) = \sum_{k=1}^K \pi_k \mathcal{N}(x | \mu_k, \Sigma_k)$$

where  $K$  is the number of mixture components,  $\pi_k$  is the mixing weight,  $k, \mu_k$  represent the component index and mean vector, respectively,  $\Sigma_k$  is the covariance matrix, and  $\mathcal{N}(x | \mu_k, \Sigma_k)$  is the Gaussian distribution. GMM was included because it can model overlapping clusters.

### 3.6. Agglomerative Clustering

Agglomerative clustering was performed using both average linkage and complete linkage. This approach starts with each observation in its own cluster and then progressively merges clusters based on distance. The study expresses the total distance between two clusters like:

$$D(C_a, C_b)$$

where  $C_a$  and  $C_b$  are two different clusters. In average linkage, the distance is the average of all pairwise distances between observations in the two clusters:

$$D_{avg}(C_a, C_b) = \frac{1}{|C_a||C_b|} \sum_{x_i \in C_a} \sum_{x_j \in C_b} d(x_i, x_j)$$

In complete linkage the distance is given by the maximum distance between any two observations in each cluster:

$$D_{complete}(C_a, C_b) = \max_{\substack{x_i \in C_a \\ x_j \in C_b}} d(x_i, x_j)$$

Agglomerative clustering was included to compare hierarchical grouping behavior with BIRCH and the probabilistic Gaussian Mixture Model.

### 3.7. DBSCAN Clustering

DBSCAN is a density-based clustering algorithm that groups observations in dense regions and can identify noise points. It uses two parameters: epsilon, the neighborhood radius, and *MinPts*, the minimum number of points required to form a dense region. A point  $x_i$  has its neighborhood defined as follows:

$$N_\epsilon(x_i) = \{x_j \in X : d(x_i, x_j) \leq \epsilon\}$$

A point is considered a core point if:

$$|N_c(x_i)| \geq MinPts$$

DBSCAN was included due to its ability to find arbitrary cluster shapes and noise. Within heterogeneous and high-dimensional insurance data, if dense areas are not clearly defined, DBSCAN can provide weak separation.

### 3.8. PCA Visualization

To depict the clustering results in a two-dimensional setting, Principal Component Analysis (PCA) was used. PCA transforms the original feature space to new orthogonal components that explain the maximum variance. This transformation may be expressed as:

$$Z = XW$$

where  $Z$  is the transformed data,  $X$  is the standardized input matrix, and  $W$  is the principal-component weight matrix. PCA was used only for visualization, not as the main clustering input.

### 3.9. Cluster Validation Metrics

Since clustering is an unsupervised technique, there is no target class for direct evaluation of accuracy. Thus, internal validation metrics were used. The Silhouette Score measures how similar an observation is to its own cluster compared to other clusters:

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

where  $a(i)$  is the average distance from observation  $i$  to the observations in the same cluster, and  $b(i)$  is the minimum average distance from observation  $i$  to another cluster. A higher score indicates better separation.

Cluster quality was evaluated using the Davies-Bouldin Index. It is defined as follows:

$$DB = \frac{1}{K} \sum_{i=1}^K \max_{j \neq i} \left( \frac{S_i + S_j}{M_{ij}} \right)$$

where  $S_i$  and  $S_j$  are the within-cluster distances for clusters  $i$  and  $j$ , respectively, while  $M_{ij}$  is the distance between their centroids. A lower Davies-Bouldin Index indicates better clustering.

Calinski-Harabasz Score was also used as an internal validation metric:

$$CH = \frac{\text{Tr}(B_k)}{\text{Tr}(W_k)} \times \frac{n - k}{k - 1}$$

where  $\text{Tr}(B_k)$  is the between-cluster dispersion,  $\text{Tr}(W_k)$  is the within-cluster dispersion, and  $n$  is the number of observations. The symbol  $k$  represents the number of clusters. A higher Calinski-Harabasz Score indicates better-defined clusters.

#### 4. Results

The results suggest that the clustering models partitioned health insurance information into meaningful risk-based categories, and BIRCH developed a practical and understandable model. Maximum mean values of charges, age, and BMI were observed in the high-risk cluster while the low-risk cluster showed the lowest values.

Table 1 shows the clustering methods used in this study. While BIRCH, Gaussian Mixture, Agglomerative Average and Agglomerative Complete were able to find 3 clusters, DBSCAN found just one dominant cluster and therefore was not considered appropriate for this dataset. The Agglomerative Average cluster, although yielding the highest Silhouette score, was not as viable because of an extremely low Calinski-Harabasz score. The final model selected was BIRCH as it provided three meaningful and interpretable groupings of risk without noise data. BIRCH was selected using the combined evidence because it achieved the highest Calinski-Harabasz Score, assigned all observations without noise, and produced more interpretable risk profiles, although Agglomerative Average performed better on the Silhouette and Davies-Bouldin measures.

**Table 1.** Clustering Methods Evaluation.

| Method                 | Clusters | Noise % | Silhouette | Davies-Bouldin | Calinski-Harabasz |
|------------------------|----------|---------|------------|----------------|-------------------|
| BIRCH                  | 3        | 0.00    | 0.0524     | 3.2789         | 80.92             |
| Gaussian Mixture       | 3        | 0.00    | 0.0415     | 3.8071         | 58.14             |
| Agglomerative Average  | 3        | 0.00    | 0.0968     | 1.6442         | 13.86             |
| Agglomerative Complete | 3        | 0.00    | 0.0376     | 3.5864         | 60.26             |
| DBSCAN                 | 1        | —       | —          | —              | —                 |

The behavioral profiles of all clustering methods involved in the study are presented in Table 2. BIRCH was the best-performing algorithm as it created three large insurance-risk subtypes. Both the Gaussian Mixture and Agglomerative techniques also returned three clusters but with a much

less clear separation. DBSCAN produced one dominant cluster, indicating that the dataset did not contain clearly density-separated groups.

**Table 2.** Cluster Profile Summary.

| Clustering Method      | Cluster Structure | Main Profile Pattern                        | Interpretation                                                              |
|------------------------|-------------------|---------------------------------------------|-----------------------------------------------------------------------------|
| BIRCH                  | 3 clusters        | Low-risk, medium-risk, and high-risk groups | Produced the most practical and interpretable risk segmentation             |
| Gaussian Mixture       | 3 clusters        | Overlapping probabilistic groups            | Useful for soft grouping, but cluster separation was weaker than BIRCH      |
| Agglomerative Average  | 3 clusters        | Hierarchical groups with uneven structure   | Showed some separation, but one cluster dominated the structure             |
| Agglomerative Complete | 3 clusters        | More separated hierarchical groups          | Better than average linkage, but still less practical than BIRCH            |
| DBSCAN                 | 1 main cluster    | Weak density-based separation               | Not suitable because it did not clearly divide the dataset into risk groups |

Figure 2 shows the number of clusters that these five methods generate. The algorithms BIRCH, Gaussian Mixture, Agglomerative Average and Agglomerative Complete returned three clusters which supported the intended interpretation of low-, medium- and high-risk. In contrast, DBSCAN created a single cluster meaning it was not able to separate out individual density-separated groups in this dataset.

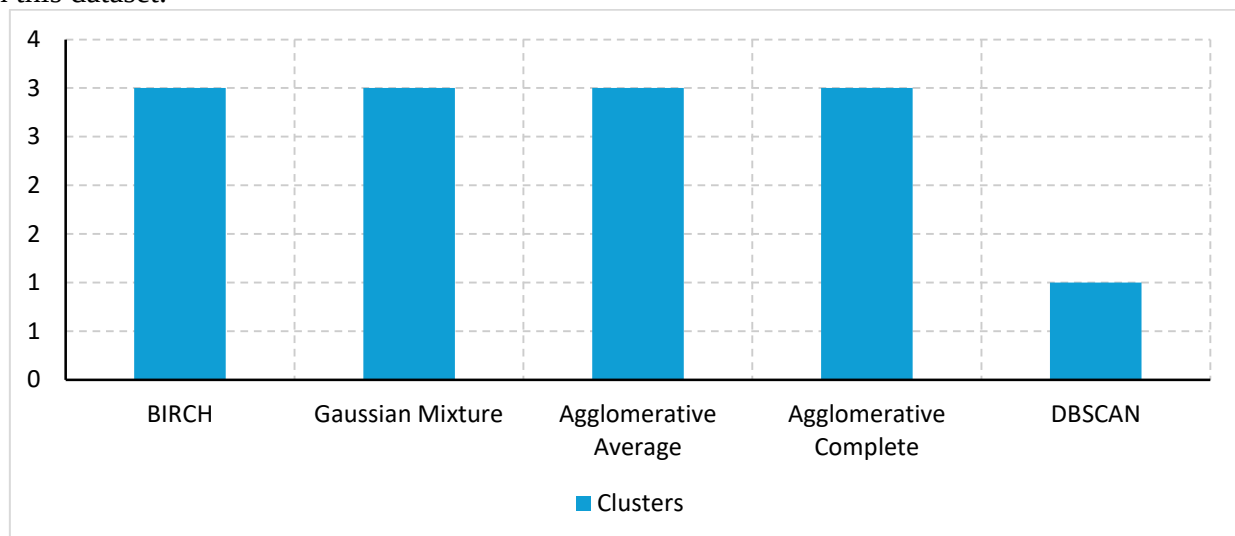


**Figure 2.** Number of Clusters Generated by Each Clustering Method.

The Davies–Bouldin Index values, calculated based on the clustering methods used in the health insurance risk assessment are depicted in Figure 3. Lower Davies–Bouldin scores indicate better separation and less overlap between the clusters. Agglomerative Average achieved the lowest value and therefore the strongest separation according to this metric, while the other methods showed higher values. DBSCAN is not represented because the Davies-Bouldin Index cannot be calculated for a single cluster.

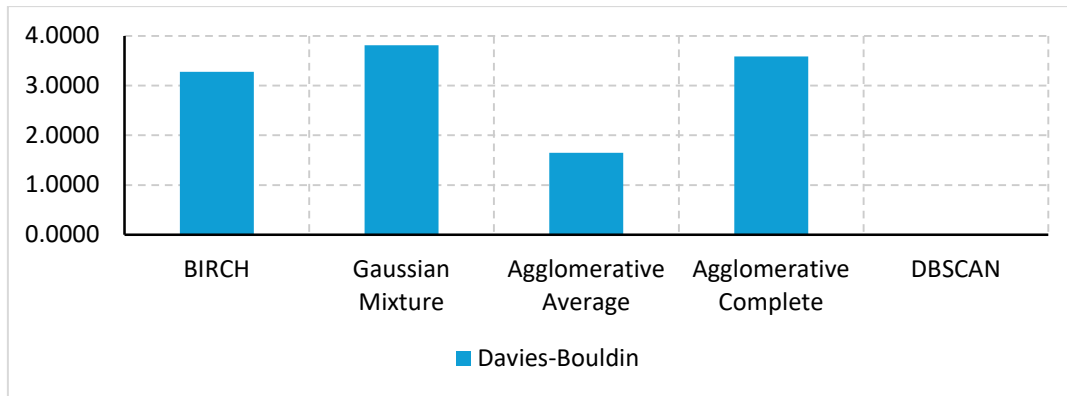


Figure 3. Davies–Bouldin Index Comparison Across Clustering Methods.

The Silhouette Score values for the clustering methods applied on this health insurance data are shown in figure 4. The Silhouette Score measures how close an observation is to the neighboring clusters in a nutshell. Higher values indicate better separation between clusters. BIRCH achieved good performance with second higher Silhouette Score following Agglomerative Average, then Gaussian Mixture and followed by Agglomerative Complete. However, all the values are relatively small meaning that we obtain overlapping cluster structure from not so well separated data. As DBSCAN produces only a single dominant cluster, the Silhouette Score cannot be calculated. The low Silhouette Scores indicate substantial overlap and weak separation; therefore, the clusters should be interpreted as exploratory risk profiles rather than fully separated actuarial classes.

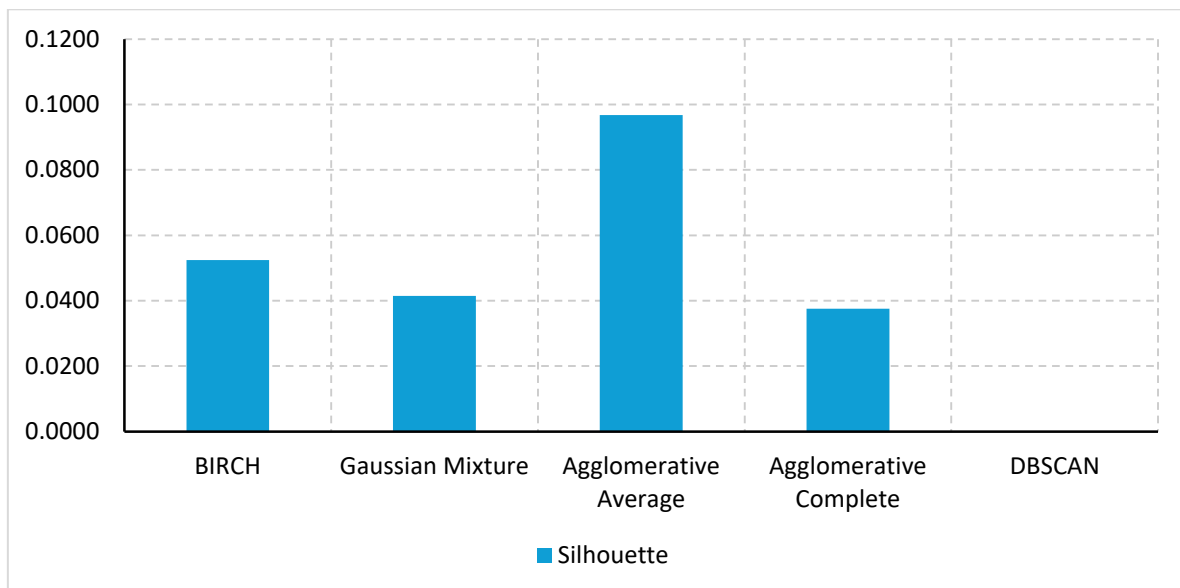


Figure 4. Silhouette Score Comparison Across Clustering Methods.

Figure 5 compares the Calinski–Harabasz Score for the five clustering methods used to evaluate insurance risk in health insurance. A larger Calinski–Harabasz Score indicates better separability between clusters and greater internal interconnectivity. BIRCH scored the highest, followed by Agglomerative Complete and Gaussian Mixture showing that in terms of this measure BIRCH resulted in the strongest overall cluster structure. The lowest score was achieved by Agglomerative Average, while DBSCAN gave no meaningful result since a single dominant cluster formed.

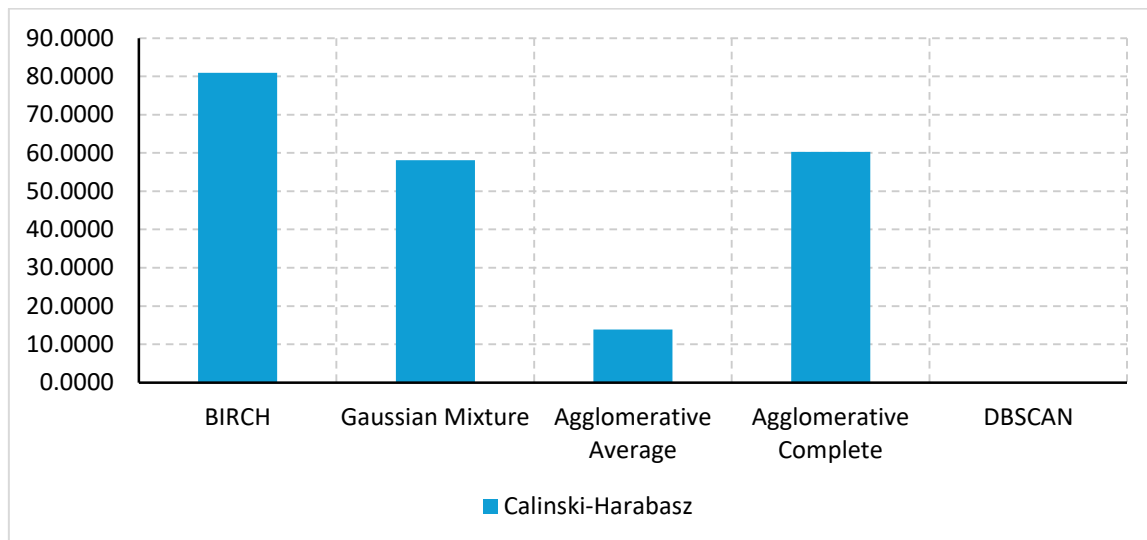


Figure 5. Calinski–Harabasz Score Comparison Across Clustering Methods.

The PCA-derived visualization from the five clustering methods run on the health insurance dataset is visualized in Figure 6. Although substantial overlap is visible among the clusters, BIRCH and Gaussian Mixture have the most discrete multi-cluster distributions among these approaches. Agglomerative Average assigns most observations to one dominant cluster, indicating poor separation and Agglomerative Complete provides better visual separation but is still highly overlapped. DBSCAN creates one dominant cluster, showing that the data do not have a density-separated structure suitable for this method. The data suggest that BIRCH is the better clustering method, as it leads to more balanced and interpretable risk groups.

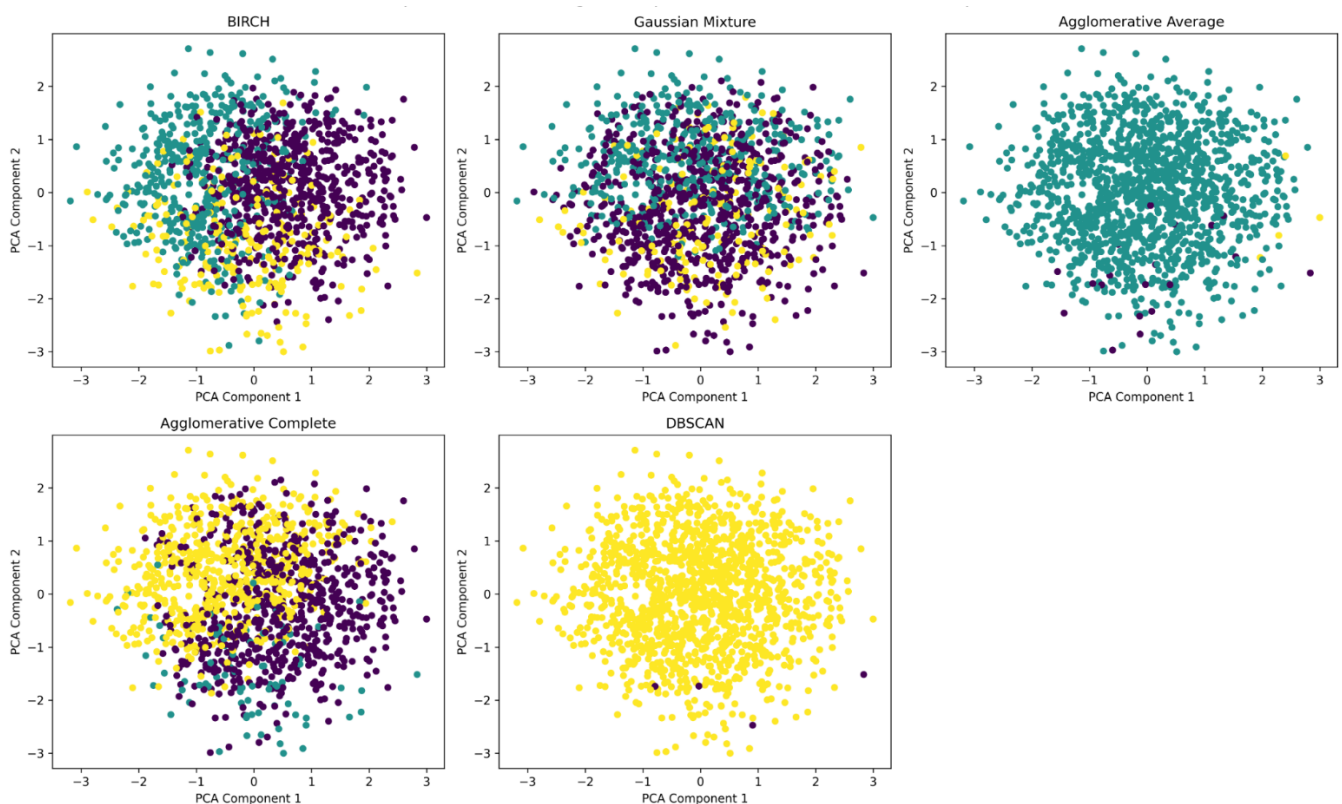


Figure 6. PCA-Based Comparison of Five Clustering Techniques for Health Insurance Risk Groups.

Figure 7 shows the comparison of Confusion matrix of the Five Clustering Approaches on the health insurance dataset. The matrices compare the actual risk groups, based on tertiles of insurance-

charge, with the predicted clusters generated by each of the unsupervised methods. BIRCH more evenly distributes over the three clusters while Gaussian Mixture defines part of the high-risk group but also has specific overlaps between the medium- and low-risk groups. Agglomerative Average and DBSCAN tend to assign observations to only one large, prevailing cluster, indicating inadequate identification of the risk-group. Agglomerative Complete provides more range than Agglomerative Average but still has a significant amount of overlap. Charges were not used to construct the clusters and were used only for external interpretation. Therefore, these matrices are descriptive comparisons rather than supervised confusion matrices.

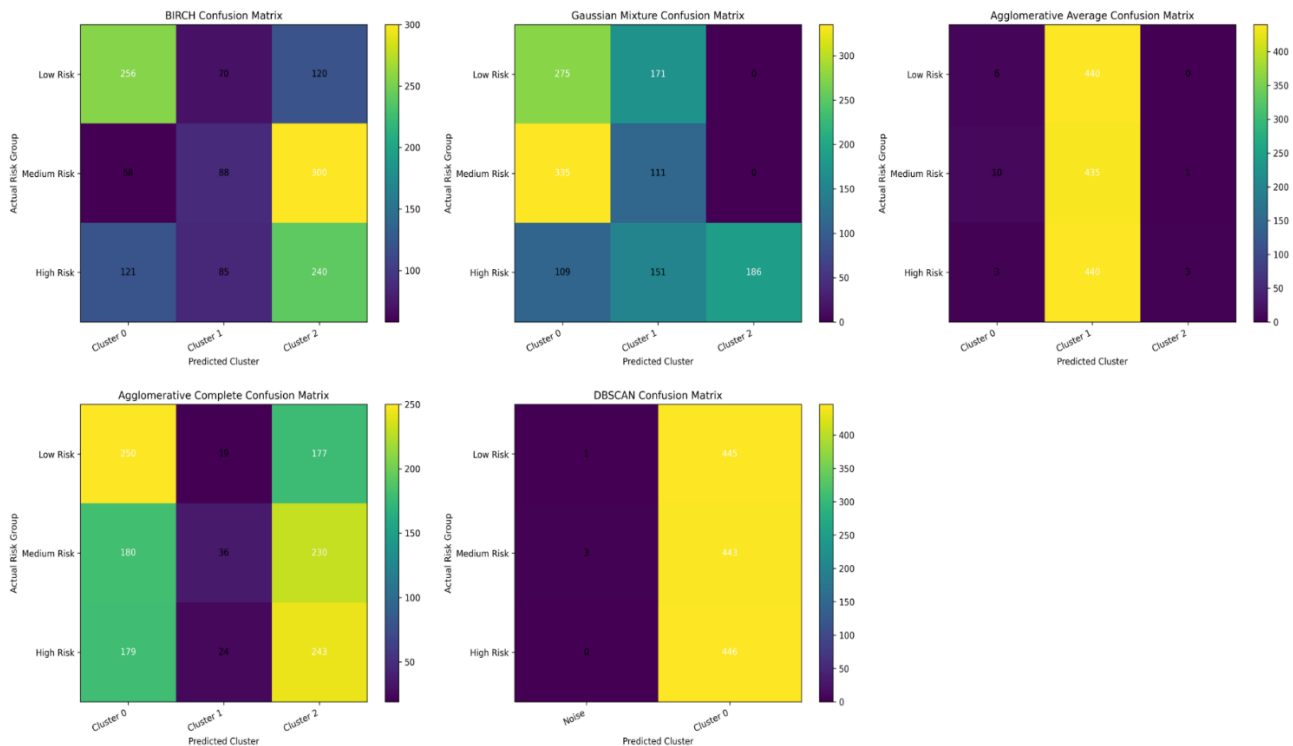


Figure 7. Confusion Matrix Comparison for Health Insurance Clustering Methods.

#### 4. Study Limitations

The study used one public dataset containing 1,338 records, which limits generalizability. The low Silhouette Scores show that the clusters overlap, and the results may vary with different preprocessing choices or model parameters. Independent or longitudinal datasets are required for further validation.

#### 5. Conclusion

This study proposed an unsupervised machine learning approach for clustering health insurance cost and risk data. The dataset was processed through several steps including data exploration, handling missing values, encoding categorical variables and feature scaling. These steps allowed the numerical and categorical variables to be used in the clustering analysis.

Various clustering methods were implemented and analyzed including BIRCH, Gaussian Mixture Model, Agglomerative Average, Agglomerative Complete, DBSCAN. The results indicated that BIRCH determined the best clustering structure that was most logical and interpretable on this dataset, as it naturally divided policyholders into three meaningful risk categories for the purpose of providing low, medium, and high-risk individuals. The final clusters were mainly distinguished by insurance charges, age, BMI, and other health-related characteristics.

The mean insurance charges were highest for the high-risk group, who had the highest average age and BMI suggesting that these two factors are significant contributors to risk of incurring high insurance costs. Conversely, the low-risk group had significantly reduced charges and more favorable profile characteristics. We found that annual income differed across clusters, but was not the core determinant of the risk level.

The results support the conclusion that clustering techniques can be beneficial to the pattern identification in health insurance data. The suggested framework is capable of assisting with insurance risk segmentation, profiling of the policyholders and analysis of healthcare cost analysis. The study can be improved in the future by using larger cohorts, including more clinical variables, applying advanced clustering optimization methods, and comparing unsupervised findings to supervised prediction models for stronger validation.

#### 4. References

1. S. Shouri, M. De la Sen, and M. E. Gordji, "Designing a smart health insurance pricing system: Integrating XGBoost and repeated Nash equilibrium in a sustainable, data-driven framework," *Information*, vol. 16, no. 9, Art. no. 733, 2025, doi: 10.3390/info16090733.
2. M. Kapse, V. Sharma, R. Vidhale, and V. Vellanki, "Customization of health insurance premiums using machine learning and explainable AI," *International Journal of Information Management Data Insights*, vol. 5, no. 1, Art. no. 100328, Jun. 2025, doi: 10.1016/j.jjime.2025.100328.
3. Scikit-learn Developers, "Clustering," *Scikit-learn User Guide*, 2026. Accessed: May 8, 2026. [Online]. Available: <https://scikit-learn.org/stable/modules/clustering.html>
4. A. Aljohani, "Optimizing patient stratification in healthcare: A comparative analysis of clustering algorithms for EHR data," *International Journal of Computational Intelligence Systems*, vol. 17, Art. no. 173, 2024, doi: 10.1007/s44196-024-00568-8.
5. A. Trezza, A. Visibelli, B. Roncaglia, O. Spiga, and A. Santucci, "Unsupervised learning in precision medicine: Unlocking personalized healthcare through AI," *Applied Sciences*, vol. 14, no. 20, Art. no. 9305, 2024, doi: 10.3390/app14209305.
6. Scikit-learn Developers, "Birch," *Scikit-learn API Reference*, 2026. Accessed: May 8, 2026. [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.Birch.html>
7. Scikit-learn Developers, "GaussianMixture," *Scikit-learn API Reference*, 2026. Accessed: May 8, 2026. [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html>
8. Scikit-learn Developers, "AgglomerativeClustering," *Scikit-learn API Reference*, 2026. Accessed: May 8, 2026. [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html>
9. Scikit-learn Developers, "DBSCAN," *Scikit-learn API Reference*, 2026. Accessed: May 8, 2026. [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html>
10. Scikit-learn Developers, "Clustering performance evaluation," *Scikit-learn User Guide*, 2026. Accessed: May 8, 2026. [Online]. Available: <https://scikit-learn.org/stable/modules/clustering.html>
11. S. B. Bhattacharya, G. C. Castignani, L. Masello, and B. S. Sheehan, "AI revolution in insurance: Bridging research and reality," *Frontiers in Artificial Intelligence*, vol. 8, Art. no. 1568266, 2025, doi: 10.3389/frai.2025.1568266.
12. A. du Preez, S. Bhattacharya, P. Beling, and E. Bowen, "Fraud detection in healthcare claims using machine learning: A systematic review," *Artificial Intelligence in Medicine*, vol. 160, Art. no. 103061, Feb. 2025, doi: 10.1016/j.artmed.2024.103061.
13. E. A. Seyam, "Predicting high-cost healthcare utilization using machine learning: A multi-service risk stratification analysis in EU-based private group health insurance," *Risks*, vol. 13, no. 7, Art. no. 133, 2025, doi: 10.3390/risks13070133.

14. Y. Bu et al., "Enhancing predictions of health insurance overspending risk through hospital departmental performance indicators," *Biomolecules and Biomedicine*, vol. 25, no. 10, pp. 2269–2280, 2025, doi: 10.17305/bb.2025.12051.
15. C. Jamotton, D. Hainaut, and T. Hames, "Insurance analytics with clustering techniques," *Risks*, vol. 12, no. 9, Art. no. 141, 2024, doi: 10.3390/risks12090141.
16. F. Alshrouf, "Improving BIRCH hierarchical clustering algorithms for enhanced partitioning of medical data," *Cluster Computing*, vol. 28, no. 9, 2025, doi: 10.1007/s10586-025-05270-8.
17. E. Onojete et al., "Enhancing disease clustering through symptom-based analysis and large language model interpretations," *Scientific Reports*, vol. 15, Art. no. 36651, 2025, doi: 10.1038/s41598-025-20382-2.
18. B. Rawat, H. Pant, and A. Bist, "Clustering medical conditions in patient records using unsupervised learning techniques: A comparative study," *Biomedical and Pharmacology Journal*, vol. 18, no. 3, 2025, doi: 10.13005/bpj/3236.
19. M. Kossakov, A. Mukasheva, G. Balbayev, S. Seidazimov, D. Mukammejanova, and M. Sydybayeva, "Quantitative comparison of machine learning clustering methods for tuberculosis data analysis," *Engineering Proceedings*, vol. 60, no. 1, Art. no. 20, 2024, doi: 10.3390/engproc2024060020.
20. S. E. Abhadiomhen, E. O. Nzeakor, and K. Oyibo, "Health risk assessment using machine learning: Systematic review," *Electronics*, vol. 13, no. 22, Art. no. 4405, 2024, doi: 10.3390/electronics13224405.
21. D. Chicco, A. Campagner, A. Spagnolo, and G. Jurman, "The Silhouette coefficient and the Davies-Bouldin index are more informative than Dunn index, Calinski-Harabasz index, Shannon entropy, and Gap statistic," *PeerJ Computer Science*, Art. no. e3309, 2025, doi: 10.7717/peerj-cs.3309.
22. W.-C. Yang, J.-P. Lai, Y.-H. Liu, Y.-L. Lin, H.-P. Hou, and P.-F. Pai, "Using medical data and clustering techniques for a smart healthcare system," *Electronics*, vol. 13, no. 1, Art. no. 140, 2024, doi: 10.3390/electronics13010140.
23. T. Beaney et al., "Clustering in multiple long-term conditions: Methodological and translational challenges and solutions," *Journal of the Royal Society of Medicine*, vol. 119, no. 1, pp. 5–9, Jan. 2026, doi: 10.1177/01410768251395387.
24. M. Zonayed, R. Tasnim, S. S. Jhara, M. A. Mimona, M. R. Hussein, M. H. Mobarak, and U. Salma, "Machine learning and IoT in healthcare: Recent advancements, challenges & future direction," *Advances in Biomarker Sciences and Technology*, vol. 7, pp. 335–364, 2025, doi: 10.1016/j.abst.2025.08.006.
25. M. Jawad, "Health Insurance Cost & Risk Dataset," *Kaggle*, 2025. [Online]. Available: Kaggle, Health Insurance Cost & Risk Dataset. Accessed: May 23, 2026.