Proceedings of 12th IOE Graduate Conference
Peer Reviewed
ISSN: 2350-8914 (Online), 2350-8906 (Print)
Year: 2022 Month: October Volume: 12

Risk Stratification in Healthcare Data using Clustering Algorithms

Saurav Prajapati [a], Arun kumar Timalsina [b]

[a, b] *Department of Electronics and Computer Engineering, Pulchowk Campus, IOE, Tribhuvan University, Nepal*
[a] 075mscsk018.saurav@pcampus.edu.np, [b] t.arun@ioe.edu.np

Abstract

Risk stratification in healthcare is about grouping the patients based on their historical information and risk of spending higher amount in near future. There are different methods for risk stratification in healthcare. Simply using only diagnosis for grouping is a simple approach but that requires extensive domain knowledge. Machine learning algorithms are also used for risk stratification. Regression, classification and clustering can be used for risk stratification. In this research, clustering is used for risk stratification. 2 data periods are used: 1 year and 3 years, and implemented 3 different clustering algorithms: k-means, DBSCAN and mean-shift. For data transformation, scaling and PCA are used. Evaluated results after each data transformation in each of clustering algorithms and compared the results. For internal evaluation silhouette score, C-H index and D-B index are used. Among them, silhouette score is used for comparing the results with one another. The best result is in k-means clustering with data normalization in 1 year dataset. The best silhouette score is 0.689 with k=8.

Keywords

Clustering, DBSCAN, K-Means, Mean-Shift, Risk Stratification

1. Introduction

Healthcare cost is increasing at an alarming rate all over the world. Increasing health care cost is one of the most challenging problem. Different approaches have been made for mitigating this problem. One of the solutions that can be done is identify the highly risky members and apply preventive measures for them. In a study done in Medicare population in US, McWillians and Schwatrz [1] found that 17 percent of the population incurred 75 percent of all costs. If we can apply risk stratification among the patients, identify high risk members then it would help to reduce future healthcare cost. Traditionally for risk stratification, grouping of the patients were done based on their diagnosis. It would require medical expertise and a lot more knowledge. Diagnostic Cost Groups (DCG) model [2] is one such example. This is where application of machine learning come into use.

Regression, classification, and clustering can be used for this process. Using regression future amount is predicted on individual basis and stratify individuals according to the predicted cost. Different ML algorithms can be used for this such as linear regression, decision tree regression, deep learning, ensemble methods, etc. Using this approach, following formula is widely used for calculating individual risk score.

Individual risk score:

$$\frac{predicted\ individual\ amount}{Average\ amount\ across\ the\ population} \tag{1}$$

Classification can also be done for risk stratification. It is simpler approach. First, we divide the data into different classes based on future costs and we apply classification methods for this. Different classification methods such as logistic regression, decision tree classification, deep learning methods, and ensemble methods.

Clustering can also be done for risk stratification. It is used when data is not labelled. In this approach, similar kind of patients are grouped together. Different clustering algorithms can be used such as k-means, DBSCAN, mean-shift, etc.

2. Literature Review

Before working on, some of the previous researches are studied on health care risk stratification.

Jose et. al.[3] used supervised machine learning techniques for predicting high-cost claimants. Those who have annual cost of greater than 250,000 were considered high-cost claimants. They tried to predict those whose expenditure would exceed 250,000 in next 12 months. The reporting period is 1 year and the prediction period is 1 year. They had the highest accuracy with auc-roc score of 91.25 percent using Gradient boosted tree classifier.

Bertimas et. al.[4] used 2 years data to predict the amount in 3 rd year. They created different cost buckets: ¡$3200, $3200-$8000, $8000-$18000, $18000-$50000 and greater than $50000. They used different algorithms: regression, classification and clustering.

Sushmita et. al.[5] used ML algorithms to predict individual health care cost on individual basis. They used regression tree, M5 tree and random forest. They used 4 different scenarios, using 3 months data to predict following 9 months data. Using 6 months data to predict following 6 months data, using 9 months data to predict following 3 months data and using 12 months data to predict following 12 months data. In their study, they have found that M5 tree had best result for predicting future cost.

Morid et. al.[6] used time series data for predicting future health care cost. They used 2 years as reporting period and 1 year as prediction period. They used the CNN for training the model. They used temporal data for each month. They have Mean Absolute percentage Error of 1.6 percentage.

4Xianlong et. al.[7] used unsupervised method for risk stratification and analysis. They proposed a new framework transformer based multimodal auto encoder(TMAE) for data embedding and used clustering for risk stratification of the data. They have used more than 600,000 members. Their framework for embedding the data contains two different processes. One is encoding and another is decoding. In both of the process, they used classical transformer architecture. The use of this auto encoder is to capture temporal information. They tested the result using some baseline embedding techniques and found that their approach has the highest accuracy.

The traditional way of analyzing risk in healthcare data is using supervised machine learning algorithms: classification [3][4] and regression [5] [6]. These supervised algorithms are used to predict unseen future events and the members are stratified accordingly. For example in regression, looking back past 1 year data the next 1 year healthcare cost is predicted. It is hard to predict future events perfectly. This is the reason why supervised algorithms have less accuracy. There are studies done in unsupervised learning [7] where clustering algorithms are used to cluster members within a certain groups for example diabetes. Clustering algorithms can be used to stratify members using their healthcare data. Since it uses past historical information only, it should be more accurate.

In this research, we have tried to show that clustering can be used for risk analysis in healthcare data. Therefore based on above discussions unsupervised learning algorithms: k-means, DBSCAN and mean-shift are selected for clustering purpose.

3. Methodology

In this research, clustering algorithms are used for risk stratification purpose and compared the results among them. The process includes data collection, analysis, feature engineering, implementation of algorithms and evaluation.

3.1 Data

CMS 2008-2010 Data Entrepreneurs' Synthetic public use file data, sample 1 [8] from Centers for Medicare and Medicaid services(CMS) is used in this research. Following are the details of data that is used.

Table 1: Beneficiary claims data description

Year	Beneficiary claims
2008	116,352
2009	114,538
2010	112,754

Besides beneficiary claims, we used following different files.

Table 2: Data description

Inpatient	Outpatient	Prescription
66,773	790,790	5,552,421

3.2 Data preparation

The data contains 3 years of data. 2 different datasets are prepared: 1 year data and 3 years data. Both datasets are tested separately. There are total 21 attributes. All the attributes used and its description are given in Table 3.

Table 3: Attributes

Attribute	Description
Sex	Gender of the person
Age	Age of the person
SP_ALZHDMTA	Chronic condition Alzheimer
SP_CHF	Chronic condition Heart failure
SP_CHRNKIDN	Chronic condition chronic kidney disease
SP_CNCR	Chronic condition Cancer
SP_COPD	Chronic condition COPD
SP_DEPRESSN	Chronic condition Depression
SP_DIABETES	Chronic condition Diabetes
SP_ISCHMCHT	Chronic condition ischemic heart disease
SP_OSTEOPRS	Chronic condition Osteoporosis
SP_RA_OA	Chronic condition rheumatoid arthritis and osteoarthritis
SP_STRKETIA	Chronic condition Stroke/transient ischemic attack
Inpatient_amount	Total inpatient amount
Inpatient_claims_count	Total inpatient claims count
Outpatient_amount	Total outpatient amount
Outpatient_claims_count	Total outpatient claims count
Pharmacy_amount	Total pharmacy amount
Pharmacy_claims_count	Total pharmacy claims count
Total_days_of_supply	Total days of supply of drugs for the given period
Total_drug_quantity	Total drug quantity

In 3 years dataset, it has 3 years of information and in 1 year dataset, it has 1 year of information. For 3 years data, included those members who are eligible in all 3 years and for 1 year data included members eligible in 2010. First, removed members, who died in 2010. Used 4 different files: beneficiary, inpatient, outpatient and drug event file. Since it is a temporal data, need to aggregate information. For each member, a row is created. Amounts and other continuous features are summed. For demographic and other chronic features, used information from 2010.

Three years dataset has 110,890 members and one year dataset has also 110,890 members.

3.3 Feature Engineering

2 different datasets are prepared. Each dataset contains 21 attributes extracted from eligibility, utilization, medication, expenditure and visits data. Transformed data using different techniques. Each process is explained briefly below:

3.3.1 Scaling

Distance based algorithms can be affected by range of attributes. In our dataset, amounts can range from 0 to thousands, but age can only range from 1 to about 100. In this kind of examples, higher magnitudes features can dominate low magnitude features, which may affect the result. So 2 different scaling methods are used. Compared the results between the 2 scaling methods.

Standardization/Standard scaling:

$$X' = \frac{(X - mean)}{standard\ deviation} \tag{2}$$

Here, the data is centered into mean value within a unit standard deviation.

Normalization/MinMax scaling:

$$X' = \frac{(X–Xmin)}{(Xmax - Xmin)} \tag{3}$$

Here, the value is transformed into a range of values. We converted into the range of 0 to 1.

3.3.2 Principal component analysis (PCA)

It is used to reduce the dimensionality of the attributes. In our case, need to plot the clusters. There are 21 attributes which we can not plot in 2d graph, so used PCA to reduce feature dimensionality into 2 and plot the clusters. Also evaluated the results after using PCA.

3.3.3 Elbow method

Elbow method is a technique which is used to find the optimum numbers of clusters in k-means clustering. In this method, initially a single cluster is assumed and within cluster sum of squares is calculated. Then number of clusters is increased one at a time. So, when a curve is plotted between number of clusters and within cluster sum of squares an elbow structures curve is plotted. From the elbow point, the curve would be almost parallel to X axis. So, the point where elbow point lies is the optimum number of clusters. For DBSCAN clustering, used elbow method for finding optimum epsilon(EPS).

3.4 Clustering Algorithms

Following 3 clustering algorithms are used in this research.

3.4.1 K-Means clustering

K-Means clustering is a centroid-based algorithm. It is one of the unsupervised learning algorithms for solving clustering problems. It is the centroid-based clustering. First number of clusters is decided. After that, centroid is randomly selected and each data point is assigned to its closest centroid. Then variance is calculated and a new centroid is defined for each cluster. Data points are reassigned to its nearest centroid. This process is repeated until there are unassigned data points left or no change in centroid. This algorithm is easy to use and implement.

3.4.2 DBSCAN clustering

DBSCAN is a density-based clustering algorithm. Unlike k means clustering, DBSCAN has ability to identify clusters with outliers and work with nonconvex clusters. It uses two parameters: eps which defines the neighborhood and minimum number of neighbors which defines minimum number of neighbors within eps. It can have different points such as core, border and outlier depending upon the position of data point.

3.4.3 Mean-Shift clustering

It is also a density-based clustering algorithm. It looks for stationary points at first. Then, clusters are shifted to higher density region shifting the center of cluster to mean of points. The process is repeated till no more points can be added. It is non-parametric and no need to specify number of clusters prior. The shape of cluster is not limited so it can be of any shape.

3.5 Evaluation

For evaluation of clustering, we can use following evaluation metrics:

3.5.1 Silhouette coefficient

It is calculated for each sample and ultimately average value is calculated. For each sample, two different distances are calculated: a: mean distance between a sample and all other points in the same cluster. b: mean distance between a sample and all other points in the nearest cluster.

For each sample, silhouette coefficient is:

$$silhouettecoefficient = \frac{(b–a)}{max(a,b)} \tag{4}$$

Score can be between -1 and 1. Higher the score better it is. 1 means the clusters are completely separated.

3.5.2 Calinski-Harabasz Index (C-H index)

It is the ratio of betwee n-cluster dispersion and inter-cluster dispersion. Higher the score better it is.

If the data is E with size n_E clustered into k clusters, then Calinski-Harabasz score is defined by following equation:

$$s = \frac{tr(B_k)}{tr(W_k)} \times \frac{n_E - k}{k-1} \tag{5}$$

$tr(B_k)$ = trace of between group dispersion matrix
$tr(W_k)$ = trace of within cluster dispersion matrix

3.5.3 Davies-Bouldin Index (D-B index)

It signifies the similarity between cluster. Lower index refers to better separation between the clusters.

$$DB = \frac{1}{k}\sum_{i=1}^{k} maxR_{ij}, where\ i \neq j \tag{6}$$

R_{ij} = ration of sum of cluster diameter of 2 different clusters to distance between cluster centroids i and j.

4. Results and Discussion

3 different clustering algorithms are implemented. Following are the results in each of them:

4.1 K-Means clustering

There are 2 different datasets. So for each dataset, evaluation is done separately.

4.1.1 3 years dataset

At first, elbow method is used for finding out the optimum number of clusters.

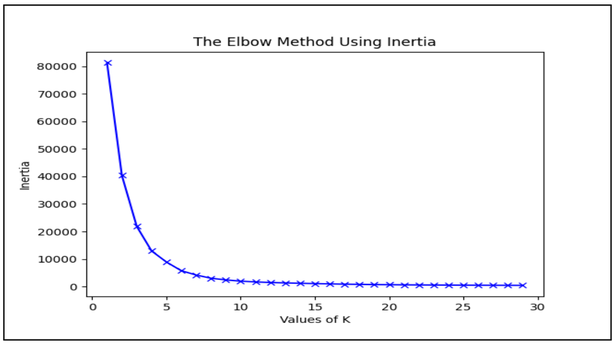

Figure 1: Elbow curve in 3 years data for optimum k

Looking at the elbow curve in figure 1, the optimum number of clusters is not quite clear. So tested the results with multiple k values : 3, 4, 5, 6, 7, 8 and 9. First applied standard scaling which is also called standardization.

Table 4: Table showing k-means results in 3 years data with standard scaling

Clusters	Silhouette score	DB score	CH score
3	0.24	2.25	17337.42
4	0.45	0.81	97589.68
5	0.44	0.81	97602.28
6	0.44	0.84	98007.50
7	0.46	0.74	104209.13
8	0.45	0.75	106291.93
9	0.44	0.75	107662.30

Using standard scaling in table 4, the best silhouette score of 0.46 is achieved, when cluster size is 7. Then minmax scaling is applied.

Table 5: Table showing k-means results in 3 years data with minmax scaling

Clusters	Silhouette score	DB score	CH score
3	0.20	1.66	23118.42
4	0.62	0.61	195269.26
5	0.63	0.54	225007.62
6	0.650	0.52	290985.77
7	0.650	0.51	338948.04
8	0.655	0.50	410786.79
9	0.649	0.51	451875.52

In table 5, the highest silhouette score is achieved when cluster size is 8. The value is 0.655. For plotting the clusters, need to apply dimensionality reduction technique. So PCA is implemented and selected top 2 components. The clusters are plotted and also applied k-means to it.

Table 6: Table showing k-means results in 3 years data with minmax scaling and PCA

Clusters	Silhouette score	DB score	CH score
8	0.655	0.50	410786.79

In table 6, after using minmax scaling and PCA the result is similar to using minmax scaling only.

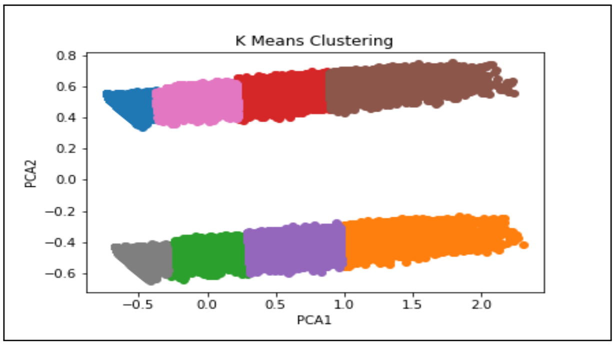

Figure 2: K-Means clustering in 3 years data with minmax scaling and PCA

The best silhouette score in 3 years data is 0.655. The clusters are in figure 2. The optimum cluster size is 8. This result is after applying minmax scaling and PCA.

4.1.2 1 year dataset

First, elbow method is used for finding out the optimum number of clusters.

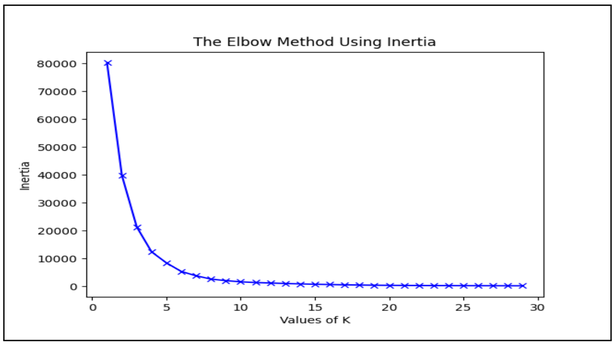

Figure 3: Elbow curve in 1 year data for optimum k

Looking at the elbow plot in figure 3, the optimum number of clusters is not quite clear. So, the results are evaluated with multiple k values : 3, 4, 5, 6, 7 , 8 and 9. First standard scaling is used.

Table 7: Table showing k-means results in 1 year data with standard scaling

Clusters	Silhouette score	DB score	CH score
3	0.24	2.23	16237.78
4	0.47	0.82	96523.32
5	0.45	0.86	97489.18
6	0.46	0.80	99636.24
7	0.46	0.77	100325.29
8	0.45	0.76	101530.14
9	0.45	0.77	104497.96

Table 7 is the results in different cluster size after applying standard scaling. The best silhouette score is 0.47. Then minmax scaling is applied.

Table 8: Table showing k-means results in 1 year data with minmax scaling

Clusters	Silhouette score	DB score	CH score
3	0.21	1.62	22852.38
4	0.64	0.59	201994.25
5	0.65	0.52	234880.80
6	0.67	0.49	308946.10
7	0.681	0.48	368744.89
8	0.689	0.47	463827.25
9	0.687	0.48	518633.15

Table 8 is the results after applying minmax scaling. Here the best results is when k=8 and silhouette score is 0.689. So for plotting the clusters, PCA is implemented. Also applied k-means clustering after PCA in table 9.

Table 9: Table showing k-means results in 1 year data with minmax scaling and PCA

Clusters	Silhouette score	DB score	CH score
8	0.689	0.47	463832.85

Using PCA after minmax scaling did not increase the result in table 9. Then clusters are plotted for k=8.

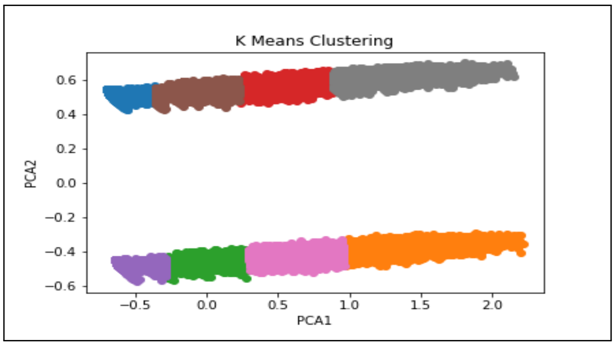

Figure 4: K-Means clustering in 1 year data with minmax scaling and PCA

So the highest silhouette score in 1 year data is 0.689 after applying minmax scaling. The clusters are in figure 4. There are 8 clusters.

4.2 DBSCAN clustering

Same as k-means, first worked with 3 years dataset and then 1 year dataset.

4.2.1 3 years dataset

Elbow method is used for finding optimum EPS value. Nearest neighbor is calculated for each data and plotted the curve. In k-means result, minmax scaling has better result than standard scaling. So first minmax scaling is applied and finally implemented PCA. After each transformation, the values changed so had to generate elbow method for finding optimum EPS after each transformation. For minimum number of data required, we used 2500.

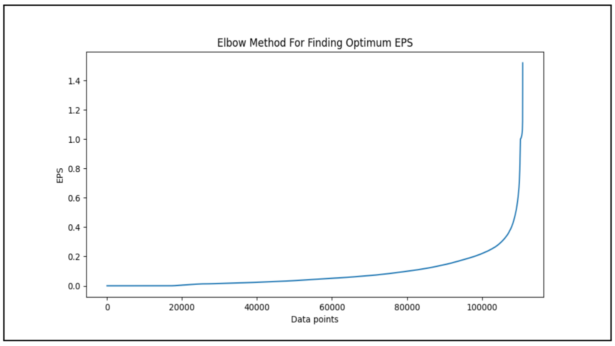

Figure 5: Elbow curve in 3 years data with minmax scaling for optimum EPS

From figure 5, the optimum EPS is around 0.4. So tested with multiple values around 0.4 EPS.

Table 10: Table showing DBSCAN results in 3 years data with minmax scaling

EPS	Silhouette score	DB score	CH score
0.2	0.09	1.55	11274.02
0.4	0.13	1.55	13782.86
0.6	0.13	1.6	10249.82
0.8	-0.16	1.49	3508.42

From table 10, the best silhouette score is 0.13 only. After that PCA is implemented. First transformed the data and plotted elbow curve for optimum EPS.

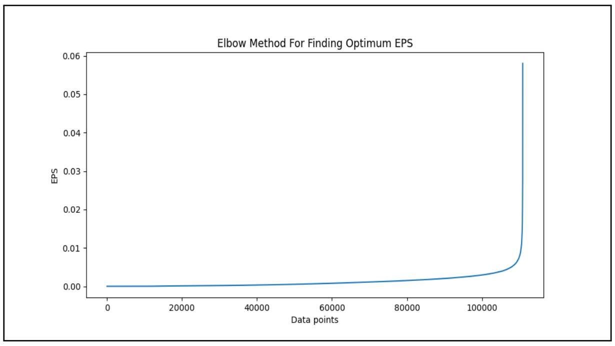

Figure 6: Elbow curve in 3 years data with minmax scaling and PCA for optimum EPS

The optimum EPS value is not clear in figure 6. So used multiple EPS values around 0.01.

Table 11 has the results when used minmax scaling and PCA. The best silhouette score after PCA is 0.27 in 3 years data. Then plotted the curve for maximum silhouette score. The plot is in figure 7.

Table 11: Table showing DBSCAN results in 3 years data with minmax scaling and PCA

EPS	Silhouette score	DB score	CH score
0.005	-0.16	0.89	8678.69
0.01	-0.001	0.83	15751.68
0.02	0.27	0.71	40758.55

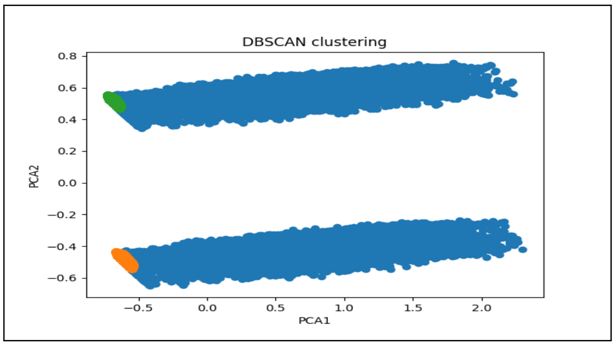

Figure 7: DBSCAN clustering in 3 years data with MinMax scaling and PCA

The highest silhouette score is 0.27 only in 3 years dataset with DBSCAN clustering. The optimum cluster size is 3 only in figure 7. It did not improve results compared to K-Means algorithm.

4.2.2 1 year dataset

For 1 year dataset, followed the same approach as in 3 years dataset. First used generated elbow curve for optimum EPS.

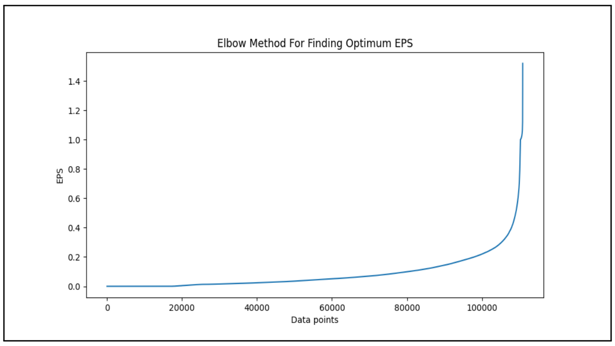

Figure 8: Elbow method in 1 year data with minmax scaling for optimum EPS

Looking at the elbow plot in figure 8, the optimum number of EPS is around 0.4. So tested with multiple values around 0.4.

Table 12 has the results when minmax scaling is

Table 12: Table showing DBSCAN results in 1 year data with minmax scaling

EPS	Silhouette score	DB score	CH score
0.2	0.12	1.5	12310.67
0.4	0.15	1.54	10405.67
0.6	0.15	1.56	10534.69
0.8	0.15	1.53	8695.82

applied in 1 year dataset. The best silhouette score is 0.15. Then used PCA.

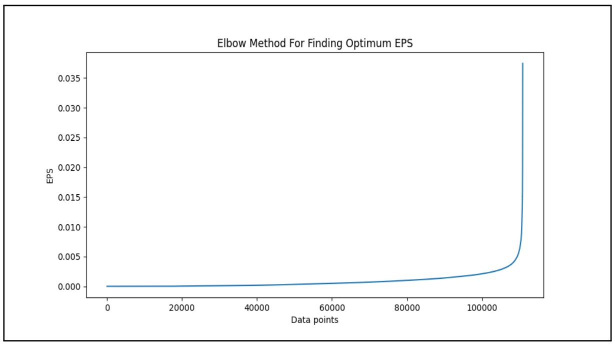

Figure 9: Elbow curve in 1 year data with minmax scaling and PCA for optimum EPS

The optimum EPS is around 0.005 in figure 9. So used multiple values around 0.005.

Table 13: Table showing DBSCAN results in 1 year data with minmax scaling and PCA

EPS	Silhouette score	DB score	CH score
0.0025	-0.24	0.93	5906.46
0.005	-0.02	0.84	15543.49
0.0075	0.16	0.76	27855.78

Table 13 is the results after applying minmax scaling and PCA. The best silhouette score is 0.16 only.

The clusters are plotted in figure 10

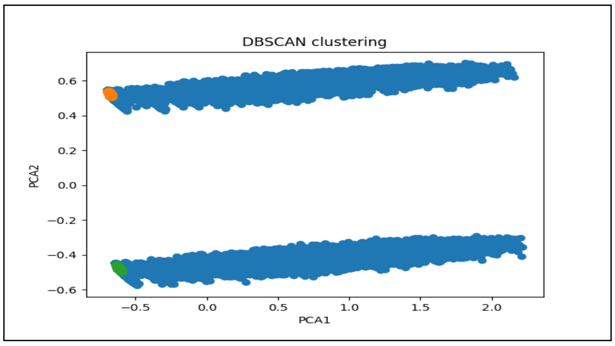

Figure 10: DBSCAN clustering in 1 year data with MinMax scaling and PCA

The best silhouette score is 0.16 in 1 year dataset. And the optimum cluster size is 3 only as in figure 10.

4.3 Mean-Shift clustering

In case of mean shift algorithm, parameters are not passed. It itself identifies the number of clusters. Same as previous, used 2 different datasets: 3 years dataset and 1 year dataset.

4.3.1 3 years dataset

Initially evaluated the result with minmax scaling. Got only one cluster so could not evaluate any result. Then applied PCA.

Table 14: Table showing Mean-Shift results in 3 years data with minmax scaling and PCA

Silhouette score	DB score	CH score
0.44	1.19	56796.26

From table 14, the silhouette score of 0.44 is achieved in meanshift clustering in 3 years data.

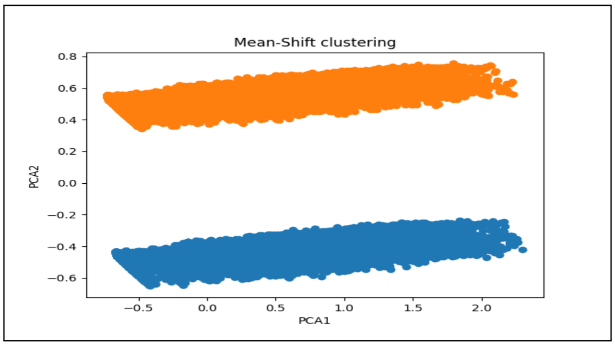

Figure 11: Mean-Shift clustering in 3 years data with minmax scaling and PCA

Th silhouette score is 0.44 in 3 years dataset which is less comapred to k-means clustering. Also the cluster size is only 2 as shown in figure 11.

4.3.2 1 year dataset

Initially the results are evaluated with minmax scaling.

Table 15: Table showing Mean-Shift results in 1 year data with minmax scaling

Silhouette score	DB score	CH score
0.35	1.57	1679.32

From table 15, the silhouette score is 0.35. Then PCA is applied.

Table 16: Table showing Mean-Shift results in 1 year data with minmax scaling and PCA

Silhouette score	DB score	CH score
0.45	1.17	58023.20

After applying PCA in table 16, silhouette score of 0.45 is achieved.

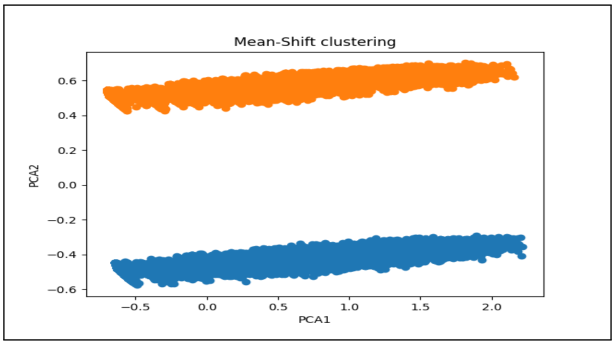

Figure 12: Mean-Shift clustering in 1 year data with minmax scaling and PCA

The best silhouette score is 0.45 in 1 year dataset. The optimum cluster size is only 2 in mean-shift algorithm as shown in figure 12.

4.4 Discussion

Clustering algorithms are implemented on healthcare data for risk stratification. With different data transformation, results are evaluated and compared with each other. The best result is silhouette score of 0.689 in k-means clustering after applying minmax scaling in 1 year dataset. The cluster size is 8. Our result is compared with the previous study on clustering in healthcare data [9], in which silhouette score is 0.3075 in hierarchical clustering. That score is surpassed using K-Means clustering in this research.

Distinct characteristics in each cluster are studied. Followings are the 8 different clusters in k-means clustering:

- All members with diabetes, high average amount and almost every has at least one pharmacy prescription.
- All male members, with no diabetes and no heart disease and very low average cost per person.
- All female members, with no diabetes and no heart disease and very low average cost per person.
- All members with chronic kidney disease and most of them have pharmacy prescriptions.
- All female members with heart disease and pharmacy prescriptions.
- All male members with heart disease and pharmacy prescriptions.
- All Members with both diabetes and heart disease and high pharmacy prescriptions.
- Other remaining members.

5. Conclusion

Three different clustering algorithms are applied for risk stratification in healthcare data. And 2 different datasets are used: 3 years and 1 year. 3 years contains 3 years of data and 1 year contains 1 year of data. Health information such as amounts, claims count, drug total days of supply, drug total quantity, chronic conditions, age and gender are included. Also different data engineering techniques are applied. Normalization has a major effect on clustering since magnitude of data directly affects the distance calculation. The best result in this research is in K-means algorithm after applying normalization. Silhouette score for comparing the results. It gives how separated the clusters are. The optimal silhouette score is 0.689 in 1 year data after applying normalization. The optimal number of clusters is 8.

6. Future Enhancements

Apart from these 3 clustering algorithms, there are other clustering algorithms which can be implemented. Internal evaluation techniques are used for evaluation. There are validation evaluation metrics as well for evaluating clusters. CMS public data is used in this research in which data are synthesized. Need to check on other datasets as well.

References

[1] J Michael McWilliams and Aaron L Schwartz. Focusing on high-cost patients: the key to addressing high costs? *The New England journal of medicine*, 376(9):807, 2017.

[2] Amy K Rosen, Susan A Loveland, Jennifer J Anderson, Cheryl S Hankin, James N Breckenridge, and Dan R Berlowitz. Diagnostic cost groups (dcgs) and concurrent utilization among patients with substance abuse disorders. *Health services research*, 37(4):1079–1103, 2002.

[3] José M Maisog, Wenhong Li, Yanchun Xu, Brian Hurley, Hetal Shah, Ryan Lemberg, Tina Borden, Stephen Bandeian, Melissa Schline, Roxanna Cross, et al. Using massive health insurance claims data to predict very high-cost claimants: a machine learning approach. *arXiv preprint arXiv:1912.13032*, 2019.

[4] Dimitris Bertsimas, Margrét V Bjarnadóttir, Michael A Kane, J Christian Kryder, Rudra Pandey, Santosh Vempala, and Grant Wang. Algorithmic prediction of health-care costs. *Operations Research*, 56(6):1382–1392, 2008.

[5] Shanu Sushmita, Stacey Newman, James Marquardt, Prabhu Ram, Viren Prasad, Martine De Cock, and Ankur Teredesai. Population cost prediction on public healthcare datasets. In *Proceedings of the 5th International Conference on Digital Health 2015*, pages 87–94, 2015.

[6] Mohammad Amin Morid, Olivia R Liu Sheng, Kensaku Kawamoto, and Samir Abdelrahman. Learning hidden patterns from patient multivariate time series data using convolutional neural networks: A case study of healthcare cost prediction. *Journal of Biomedical Informatics*, 111:103565, 2020.

[7] Xianlong Zeng, Simon Lin, and Chang Liu. Transformer-based unsupervised patient representation learning based on medical claims for risk stratification and analysis. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 1–9, 2021.

[8] Cms 2008-2010 data entrepreneurs' synthetic public use file. `https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/DE_Syn_PUF`. Accessed: 2022-08-30.

[9] M Ambigavathi and D Sridharan. Analysis of clustering algorithms in machine learning for healthcare data. In *International Conference on Advances in Computing and Data Sciences*, pages 117–128. Springer, 2020.