دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)
دانشکده علوم کامپیوتر

تمرین ششم

روش‌های پیشرفته در خوشه بندی

نگارش
ثمین مهدی پور
۹۸۳۹۰۳۹

استاد راهنما
دکتر قطعی

استادمشاور
دکتر یوسفی مهر

آذر ۱۴۰۲

چکیده

این پروژه تحلیل خوشه‌بندی را بر روی مجموعه داده بیمه انجام دادیم تا تقلب هارا شناسایی کنیم. ابتدا با استفاده از متدهای مختلف مانند KMeans، Agglomerative Clustering، و DBSCAN، داده‌ها را به خوشه‌های مختلف تقسیم کردیم. سپس با استفاده از معیارهای ارزیابی مانند Silhouette Score، نتایج خوشه‌بندی را مقایسه کردیم. همچنین، حساسیت مدل به تغییرات در پارامترها مورد بررسی قرار گرفت. در ادامه، از روش‌های کاهش ابعاد مانند PCA و t-SNE برای بصری‌سازی داده‌ها استفاده شد. این بصری‌سازی‌ها به همراه مقایسه نتایج با استفاده از معیارهای کیفیت‌سنجی، نقاط قوت و ضعف هر الگوریتم را نشان می‌دهد. در نهایت، از این نتایج برای تحلیل و ارائه جدول مقایسه‌ای از عملکرد هر الگوریتم خوشه‌بندی در مقاله استفاده می‌شود..

واژه‌های کلیدی:

خوشه بندی،، شناسایی تقلب، K-means, DB-Scan, Agglomerative Clustering

فهرست مطالب

صفحه

صفحه

فهرست اشکال

No table of figures entries found.

فصل اول

مقدمه

مقدمه

خوشه‌بندی یک فرایند مهم در تحلیل داده است که به تقسیم داده‌ها به گروه‌های مشابه یا "خوشه‌ها" بر اساس ویژگی‌های مشترکشان می‌پردازد. الگوریتم‌های خوشه‌بندی مختلف ویژگی‌های خاصی دارند که به تحلیل‌گران کمک می‌کند تا ساختار و الگوهای مهم داده‌ها را درک کنند.

۱-۱- الگوریتم های خوشه بندی

در اینجا، توضیح مختصری از سه الگوریتم خوشه‌بندی آورده شده است:

۱-۱-۱- K-Means:

- K-Means یکی از محبوب‌ترین الگوریتم‌های خوشه‌بندی است. این الگوریتم تلاش می‌کند داده‌ها را به تعدادی خوشه تقسیم کند به طوری که مرکز هر خوشه از میانگین داده‌های آن خوشه به دست آید. الگوریتم به تعداد خوشه‌ها (K) بستگی دارد.

۱-۱-۲- Agglomerative Clustering:

- Agglomerative Clustering یک الگوریتم خوشه‌بندی سلسله‌مراتبی است که از رویکرد ادغامی بهره می‌برد. در این الگوریتم، ابتدا هر داده به عنوان یک خوشه جداگانه در نظر گرفته می‌شود، سپس در هر مرحله دو خوشه به یکدیگر نزدیک‌ترین داده‌هایشان را ادغام می‌کنند تا یک سلسله‌مراتبی از خوشه‌ها ایجاد شود.

۱-۱-۳- DBSCAN (Density-Based Spatial Clustering of Applications with Noise):

- DBSCAN یک الگوریتم خوشه‌بندی مبتنی بر چگالی است که به این تفکیک می‌پردازد که داده‌هایی که در فضای چگال تر قرار دارند، به هم متصلند و به عنوان یک خوشه شناخته می‌شوند. این

الگوریتم به طور خاص برای شناسایی خوشه‌های با اشکال هندسی متفاوت و قابلیت تشخیص نویز مناسب است.

این الگوریتم‌های خوشه‌بندی در تحلیل داده‌ها به منظور شناسایی الگوها، گروه‌بندی اطلاعات مشابه، و فهم بهتر توزیع داده‌ها به کار می‌روند.

۱-۲- ارزیابی

برای ارزیابی خوشه بندی ها از معیار های زیر استفاده شده است:

۱-۲-۱- silhouette score

معیار Silhouette یک معیار ارزیابی برای خوشه‌بندی داده‌ها است که کیفیت تفکیک بین خوشه‌ها را اندازه‌گیری می‌کند. این معیار بر اساس فاصله‌ها و هم‌بستگی‌های داده‌ها درون یک خوشه و از بین خوشه‌ها محاسبه می‌شود.

فرض کنید برای هر نقطه داده، میزان همبستگی آن با داده‌های هم‌خوشه و میزان فاصله آن با داده‌های خوشه مجاورش محاسبه شود. معیار Silhouette بر اساس این میزان همبستگی و فاصله، یک امتیاز بین -۱ تا ۱ به هر نقطه اختصاص می‌دهد:

- امتیاز +۱ نشان‌دهنده این است که نقطه به درستی در خوشه‌اش قرار گرفته و از نقاط هم‌خوشه واقع در مجاورتش فاصله کمی دارد.

- امتیاز ۰ نشان‌دهنده این است که نقطه در مرز بین دو خوشه قرار گرفته است.

- امتیاز -۱ نشان‌دهنده این است که نقطه بهتر بود که در خوشه مجاور جای پیدا کرده بوده است.

معیار Silhouette بر اساس میانگین امتیازهای تمام نقاط داده محاسبه می‌شود و مقدار بالاتر از صفر نشان‌دهنده یک خوشه‌بندی خوب و متعادل است. این معیار مفید است زیرا به تفکیک و وضوح خوشه‌ها توجه می‌کند و به عنوان یک معیار جامع برای ارزیابی عملکرد الگوریتم‌های خوشه‌بندی مورد استفاده قرار می‌گیرد.

۱-۲-۲- Calinski-Harabasz

معیار Calinski-Harabasz یک معیار ارزیابی برای خوشه‌بندی داده‌ها است که بر اساس تفاوت بین داخلی و بیرونی خوشه‌ها محاسبه می‌شود. این معیار به جهت اندازه‌گیری کیفیت تفکیک میان خوشه‌ها و همچنین کیفیت یکپارچگی درون هر خوشه استفاده می‌شود.

معیار Calinski-Harabasz بر اساس واریانس داخلی (within-cluster variance) و بین خوشه‌ها (between-cluster variance) محاسبه می‌شود. این معیار به صورت زیر تعریف می‌شود:

$$\text{Calinski-Harabasz Index} = \frac{\text{Between-Cluster Variance}}{\text{Within-Cluster Variance}} \times \frac{\text{Number of Data Points} - \text{Number of Clusters}}{\text{Number of Clusters} - 1}$$

در اینجا:

- Between-Cluster Variance نشان‌دهنده واریانس بین مراکز خوشه‌ها است.
- Within-Cluster Variance نشان‌دهنده واریانس داده‌ها داخل هر خوشه است.

هدف این معیار این است که این نسبت بین داخلی و بیرونی به حداکثر برسد. به عبارت دیگر، مقدار بالاتر این معیار نشان‌دهنده خوشه‌بندی بهتر و مجزاتر است. استفاده از معیار Calinski-Harabasz می‌تواند به تحلیل‌گران کمک کند تا الگوریتم خوشه‌بندی مناسبی را انتخاب کنند و عملکرد آن را ارزیابی کنند.

۲-۲-۲- Davies-Bouldin

معیار Davies-Bouldin یک معیار ارزیابی برای خوشه‌بندی داده‌ها است که بر اساس میزان تفاوت بین هر خوشه و خوشه‌ای که با آن بیشترین تشابه را دارد، محاسبه می‌شود. این معیار به منظور ارزیابی فاصله و تفکیک بین خوشه‌ها استفاده می‌شود.

معیار Davies-Bouldin بر اساس میانگین نسبت تفاوت بین هر خوشه و خوشه‌های مجاور به بهترین خوشه (که نزدیک‌ترین خوشه مجاور باشد) محاسبه می‌شود. این معیار به صورت زیر تعریف می‌شود:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\text{Dissimilarity}(i,j)}{\text{Similarity}(i,i)} \right)$$

در اینجا:

- \(k\) تعداد خوشه‌ها است.

- Dissimilarity(i, j) نشان‌دهنده فاصله (یا اختلاف) بین خوشه i و j است.

- Similarity(i, i) نشان‌دهنده میزان تشابه داده‌های داخل خوشه i با یکدیگر است.

هدف این معیار این است که مقدار کمتری داشته باشد، که نشان‌دهنده تفکیک بهتر و مجموعه خوشه‌های متمایزتر است. استفاده از معیار Davies-Bouldin می‌تواند به تحلیل‌گران در انتخاب بهترین تعداد خوشه و ارزیابی عملکرد الگوریتم‌های خوشه‌بندی کمک کند.

۲-۲-۳- Adjusted Rand

معیار Adjusted Rand Index (ARI) یک معیار ارزیابی برای خوشه‌بندی داده‌ها است که به اندازه میزان تطابق بین یک خوشه‌بندی و یک خوشه‌بندی مرجع (که به عنوان "درست" در نظر گرفته می‌شود) می‌پردازد. این معیار به عنوان یک اندازه‌گیر نسبی از تفاوت و تطابق میان دو خوشه‌بندی عمل می‌کند.

تعریف ARI به صورت زیر است:

$$ARI = \frac{\text{Adjusted Rand Index}}{\text{Maximum Possible Adjusted Rand Index}}$$

در اینجا:

- Adjusted Rand Index میزان تطابق و تفاوت میان دو خوشه‌بندی را محاسبه می‌کند.

- Maximum Possible Adjusted Rand Index مقدار حداکثر ممکن ARI است که با فرض حالت تصادفی و بدون ساختار خوشه‌ها محاسبه می‌شود.

مقادیر ARI در بازه [۰, ۱] قرار دارند:

- ARI نزدیک به ۱ نشان‌دهنده تطابق بالا و تفاوت کم بین دو خوشه‌بندی است.

- ARI نزدیک به ۰ نشان‌دهنده تفاوت بیشتر و تطابق کم بین دو خوشه‌بندی است.

- ARI منفی نشان‌دهنده تفاوت بیشتر از حالت تصادفی است.

استفاده از ARI می‌تواند به تحلیل‌گران کمک کند تا میزان موفقیت الگوریتم خوشه‌بندی را نسبت به یک خوشه‌بندی مرجع ارزیابی کنن

فصل دوم

پیاده‌سازی

۲-۱- مجموعه داده و پیش پردازش

پس از بررسی مجموعه داده متوجه شدیم که دارای ۱۰۰۰ سطر و ۴۰ ستون است. این مجموعه داده شامل اطلاعاتی درمورد افراد تحت نظارت بیمه که متقلب دانسته شده بودند یا خیر بود. در گام بعدی داده هارا مورد بررسی قرار دادیم که دارای داده تکراری یا گم شده هستند که هیچ گونه داده تکراری یا گم شده ای یافت نشد. نهایتا چند ستون که بنظر در تعیین نتیجه نهایی مفید نبودند را حذف کردیم.

۲-۲- خوشه بندی

پس از تقسیم داده ها به نسبت ۲۰ به ۸۰ برای آزمون و آموزش و اعمال نرمالسازی شروع به اعمال الگوریتم های خوشه بندی کردیم.

۲-۲-۱- K-means

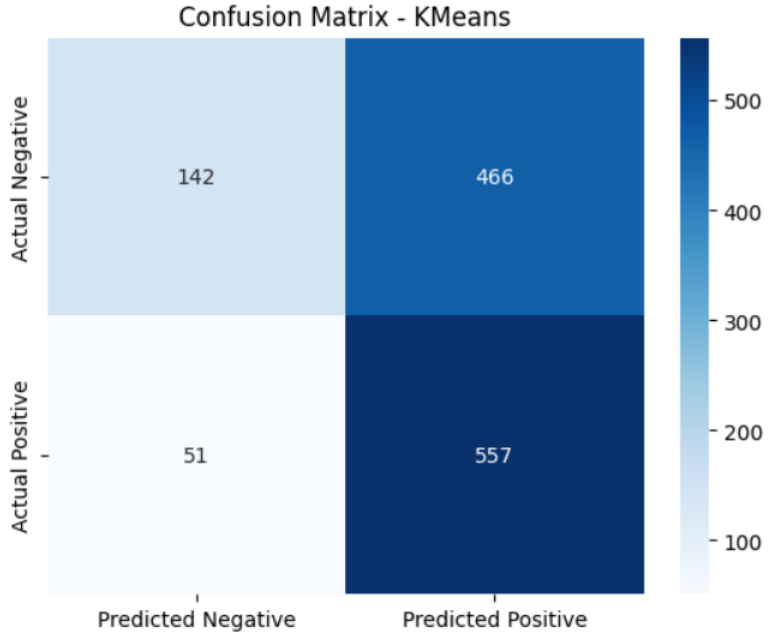
در این بخش از کد، الگوریتم K-Means با تعداد خوشه‌های ۲ و روش تصادفی با استفاده از `random_state=42` اجرا شده است. سپس از معیار Silhouette Score برای ارزیابی کیفیت تفکیک داده‌ها به خوشه‌ها استفاده شده است. مقدار به دست آمده از این معیار با استفاده از تابع `silhouette_score` از کتابخانه Scikit-Learn محاسبه شده و در نهایت با استفاده از `print` نتیجه به صورت متنی چاپ شده است.

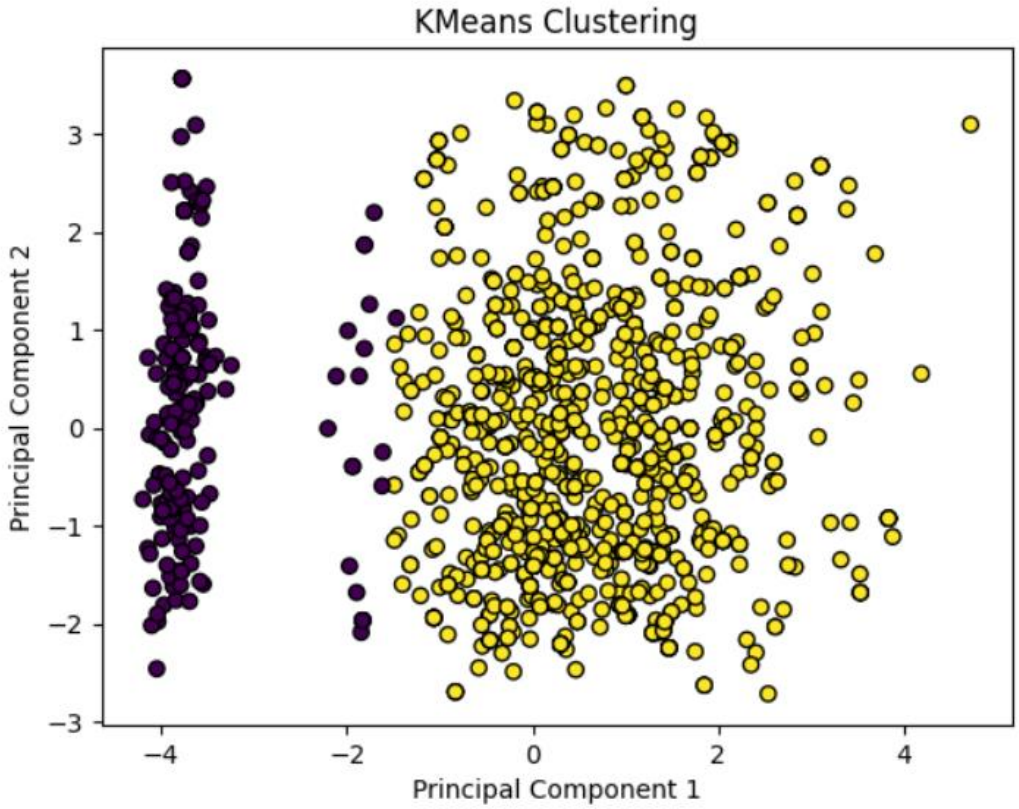
مقدار `kmeans_score` که نمایانگر Silhouette Score برای K-Means است، در اینجا به عنوان ارزیابی از کیفیت تفکیک داده‌ها به خوشه‌ها نمایش داده می‌شود. این ارزیابی در بازه [-۱, ۱] قرار دارد، که مقادیر نزدیک به ۱ نشان‌دهنده تفکیک بهتر و مجزایی از داده‌ها به خوشه‌ها هستند.

K-Means Silhouette Score: 0.15569078084268953

K-Means:

	precision	recall	f1-score	support
0	0.74	0.23	0.35	608
1	0.54	0.92	0.68	608
accuracy			0.57	1216
macro avg	0.64	0.57	0.52	1216
weighted avg	0.64	0.57	0.52	1216





۲-۲-۲ Agglomerative Clustering

در این بخش از کد، الگوریتم خوشه‌بندی سلسله مراتبی (Agglomerative Clustering) با تعداد خوشه‌های ۲ اجرا شده است. ابتدا یک نمونه از مدل Agglomerative Clustering با استفاده از `AgglomerativeClustering(n_clusters=2)` ساخته شده و سپس برچسب‌های خوشه برای داده‌های آموزش تولید شده‌اند با استفاده از `fit_predict`.

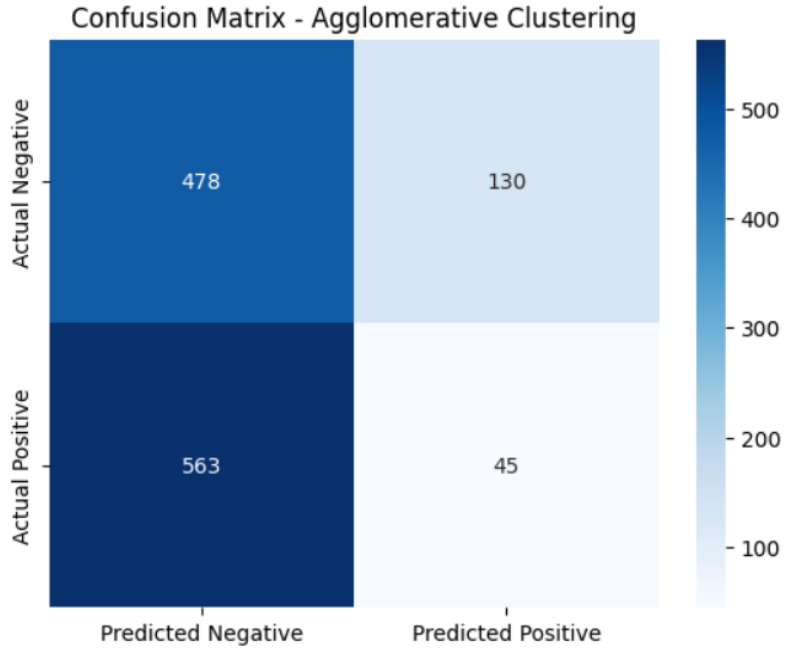
سپس از معیار Silhouette Score برای ارزیابی کیفیت تفکیک داده‌ها به خوشه‌ها استفاده شده است. مقدار به دست آمده از این معیار با استفاده از تابع `silhouette_score` از کتابخانه Scikit-Learn محاسبه شده و در نهایت با استفاده از `print` نتیجه به صورت متنی چاپ شده است.

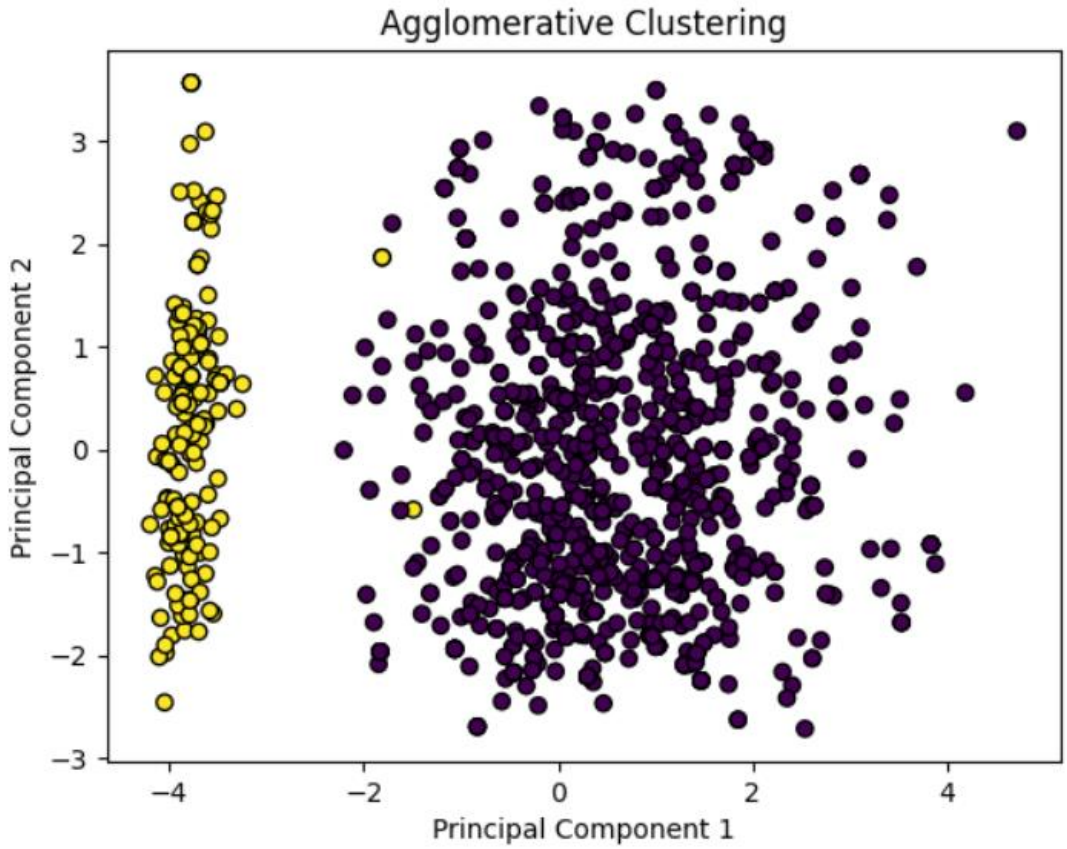
مقدار `agg_score` که نمایانگر Silhouette Score برای الگوریتم سلسله مراتبی است، نشان‌دهنده کیفیت تفکیک داده‌ها به خوشه‌ها است. این ارزیابی نیز در بازه [-۱, ۱] قرار دارد، که مقادیر نزدیک به ۱ نشان‌دهنده تفکیک بهتر داده‌ها به خوشه‌ها هستند.

Agglomerative Hierarchical Silhouette Score: 0.16083819317208767

Agglomerative Hierarchical:

	precision	recall	f1-score	support
0	0.46	0.79	0.58	608
1	0.26	0.07	0.11	608
accuracy			0.43	1216
macro avg	0.36	0.43	0.35	1216
weighted avg	0.36	0.43	0.35	1216


Confusion Matrix - Agglomerative Clustering
Actual Negative
Actual Positive
478
130
563
45
Predicted Negative
Predicted Positive
500
400
300
200
100


Agglomerative Clustering
Principal Component 2
Principal Component 1
3
2
1
0
−1
−2
−3
−4
−2
0
2
4

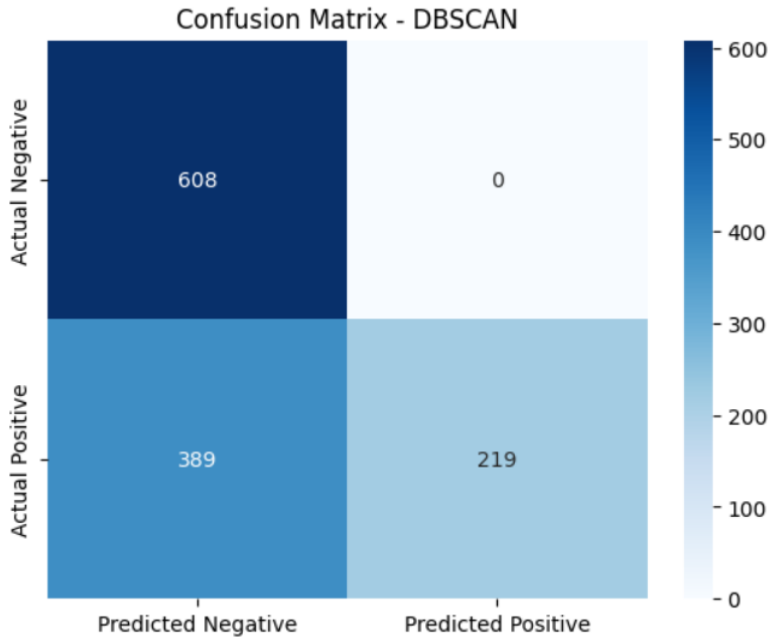
۲-۲-۳- DB-Scan

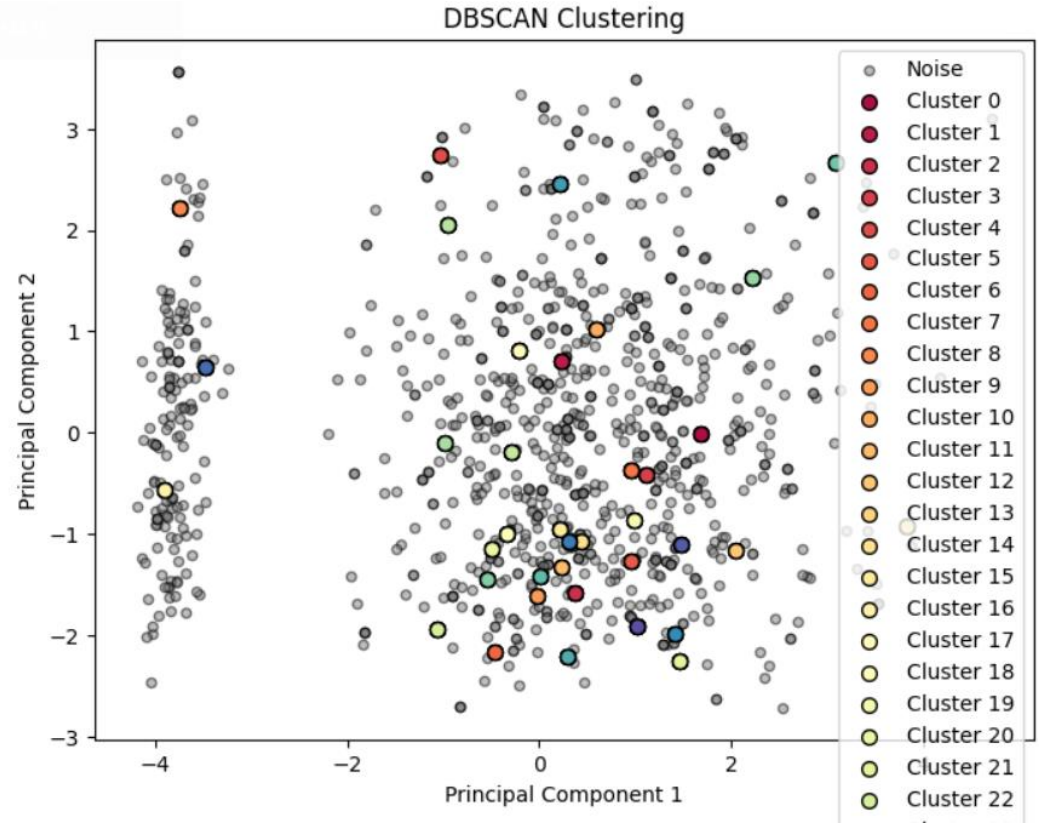
در این بخش از کد، الگوریتم DBSCAN با پارامترهای eps=0.5 و min_samples=5 اجرا شده است. ابتدا یک نمونه از مدل DBSCAN با استفاده از `DBSCAN(eps=0.5, min_samples=5)` ساخته شده و سپس برچسب‌های خوشه برای داده‌های آموزش تولید شده‌اند با استفاده از `fit_predict`.

سپس از معیار Silhouette Score برای ارزیابی کیفیت تفکیک داده‌ها به خوشه‌ها استفاده شده است. مقدار به دست آمده از این معیار با استفاده از تابع `silhouette_score` از کتابخانه Scikit-Learn محاسبه شده و در نهایت با استفاده از `print` نتیجه به صورت متنی چاپ شده است.

مقدار `dbscan_score` که نمایانگر Silhouette Score برای الگوریتم DBSCAN است، نشان‌دهنده کیفیت تفکیک داده‌ها به خوشه‌ها است. این ارزیابی نیز در بازه [-۱, ۱] قرار دارد، که مقادیر نزدیک به ۱ نشان‌دهنده تفکیک بهتر داده‌ها به خوشه‌ها هستند.

DBSCAN Silhouette Score: -0.04539766261829971





۲-۳- تحلیل حساسیت روی پارامترهای مدل‌ها

در این بخش از کد، یک تحلیل حساسیت بر روی پارامترهای مختلف الگوریتم‌های خوشه‌بندی انجام شده است. این تحلیل حساسیت بر اساس معیار Silhouette Score انجام شده و نتایج برای هر الگوریتم چاپ شده‌اند.

۲-۳-۱- K-means

برای تعداد خوشه‌ها از مقادیر ۲ تا ۹ استفاده شده است.

برای هر تعداد خوشه، الگوریتم K-Means اجرا شده و Silhouette Score برای هر حالت محاسبه شده است.

نتایج حاصل از این تحلیل بر اساس تعداد خوشه‌ها چاپ شده‌اند.

Sensitivity Analysis for K-Means:

Number of Clusters: 2, Silhouette Score: 0.1557

Number of Clusters: 3, Silhouette Score: 0.0949

Number of Clusters: 4, Silhouette Score: 0.0863

Number of Clusters: 5, Silhouette Score: 0.0769

Number of Clusters: 6, Silhouette Score: 0.0786

Number of Clusters: 7, Silhouette Score: 0.0819

Number of Clusters: 8, Silhouette Score: 0.0753

Number of Clusters: 9, Silhouette Score: 0.0766

۲-۳-۲- Agglomerative Clustering

برای انواع مختلف اتصال (linkage) از مقادیر 'ward', 'complete', 'average', 'single' استفاده شده است. برای هر نوع اتصال، الگـوریتم Agglomerative Clustering اجـرا شـده و Silhouette Score برای هر حالت محاسبه شده است.

نتایج حاصل از این تحلیل بر اساس نوع اتصال چاپ شده‌اند.

Sensitivity Analysis for Agglomerative Hierarchical:

Linkage Type: ward, Silhouette Score: 0.1608

Linkage Type: complete, Silhouette Score: 0.1548

Linkage Type: average, Silhouette Score: 0.1596

Linkage Type: single, Silhouette Score: 0.1651

۲-۳-۳- DB-Scan

برای مقادیر مختلف پارامتر eps (epsilon) از مقادیر ۰.۱, ۰.۵, ۱.۰, ۱.۵, ۲.۰, ۲.۵ استفاده شده است.

برای هر مقدار eps، الگوریتم DBSCAN اجرا شده و Silhouette Score برای هر حالت محاسبه شده است.

نتایج حاصل از این تحلیل بر اساس مقدار eps چاپ شده‌اند.

Sensitivity Analysis for DBSCAN:

Epsilon: 0.1, Silhouette Score: -0.0454

Epsilon: 0.5, Silhouette Score: -0.0454

Epsilon: 1.0, Silhouette Score: -0.0454

Epsilon: 1.5, Silhouette Score: -0.0454

Epsilon: 2.0, Silhouette Score: -0.0454

Epsilon: 2.5, Silhouette Score: -0.0454

۲-٤- مصورسازی

۲-۴-۱- K-means

این کد دو تابع به نام‌های PCA و t-SNE را برای کاهش ابعاد داده اجرا کرده و نتایج را با استفاده از الگوریتم K-Means خوشه‌بندی کرده است. سپس نمودارهای دو بعدی از داده‌ها با توجه به نتایج خوشه‌بندی رسم شده‌اند.

برای هر تابع (PCA یا t-SNE):

- ابتدا از تابع ابعاد کاهش یافته (PCA یا t-SNE) بر روی داده استفاده شده و داده‌ها به فضای دو بعدی منتقل شده‌اند.

- سپس از الگوریتم K-Means بر روی داده‌های کاهش یافته، با تعداد خوشه‌های مشخص شده (در اینجا n_clusters=2) استفاده شده است.

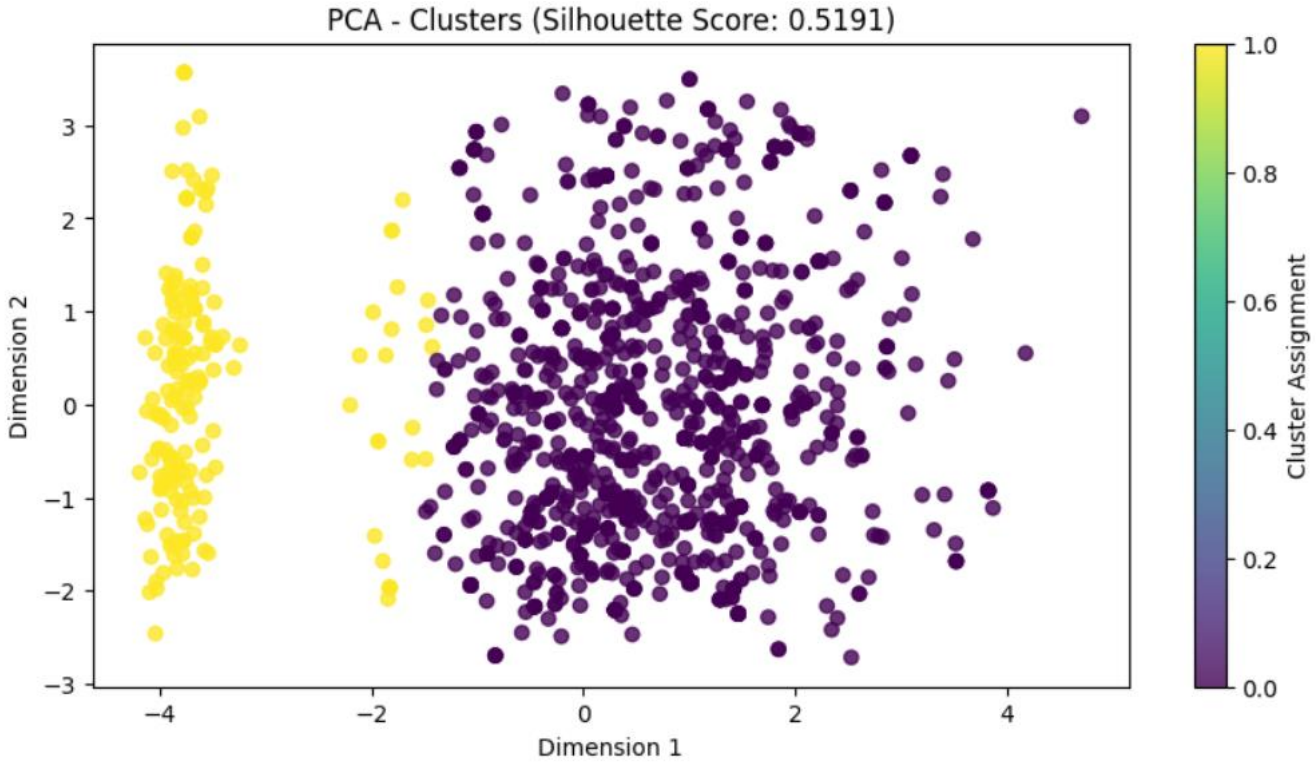
- نمودار نتیجه این خوشه‌بندی بر روی فضای دو بعدی به همراه مقدار Silhouette Score برای ارزیابی کیفیت خوشه‌بندی به تصویر کشیده شده است.

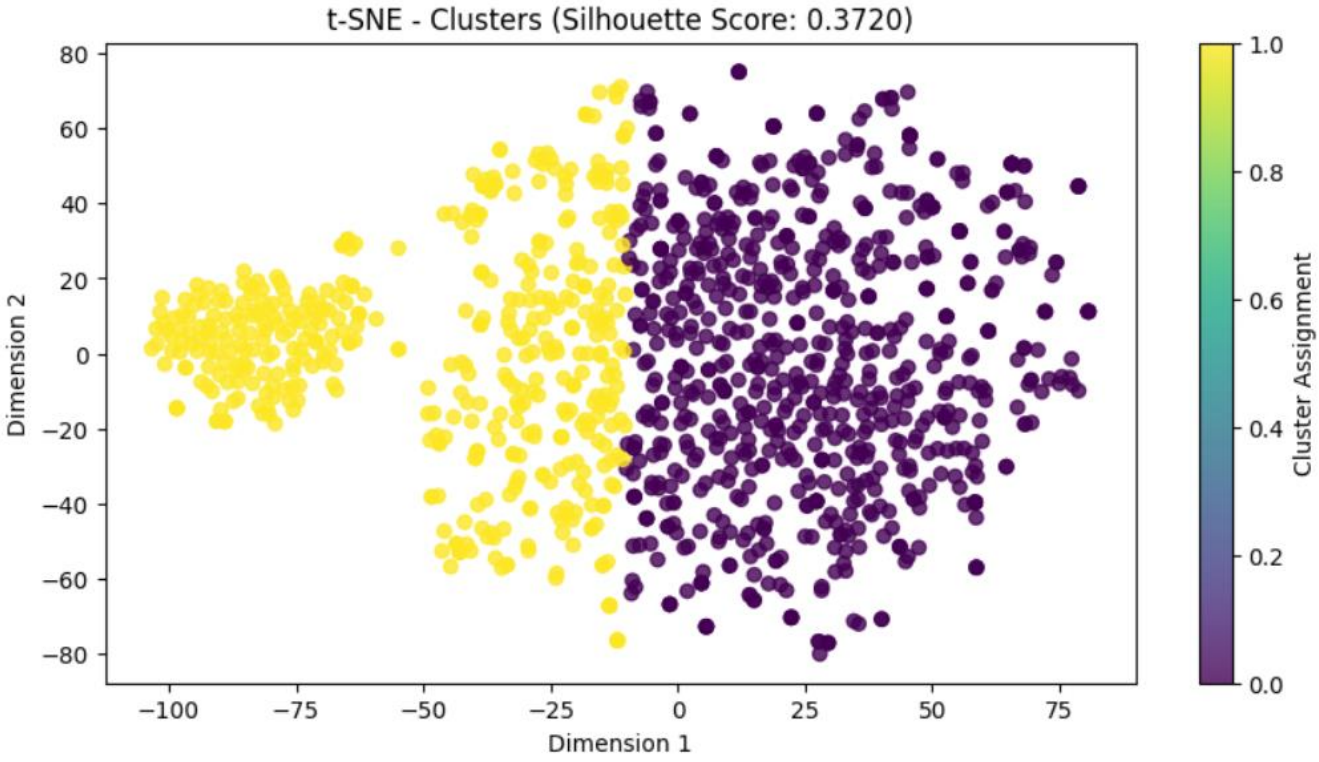
در هر نمودار:

- هر نقطه نمایانگر یک نمونه از داده است.

- رنگ هر نقطه نشان‌دهنده خوشه‌ای است که الگوریتم K-Means آن را به آن اختصاص داده است.

- مقدار Silhouette Score در عنوان نمودار ذکر شده است که نشان‌دهنده کیفیت خوشه‌بندی است. مقدار بالاتر از ۰ نشان‌دهنده تفکیک بهتر خوشه‌ها است.





Agglomerative Clustering -۲-۴-۲

این کد نیز مشابه قسمت قبل عمل می‌کند، اما از الگوریتم خوشه‌بندی Agglomerative Clustering برای تخمین خوشه‌ها استفاده می‌کند. این کد دو تابع به نام‌های PCA و t-SNE را برای کاهش ابعاد داده اجرا کرده و نتایج را با استفاده از الگوریتم Agglomerative Clustering برای خوشه‌بندی کرده است. سپس نمودارهای دو بعدی از داده‌ها با توجه به نتایج خوشه‌بندی رسم شده‌اند.

برای هر تابع (PCA یا t-SNE):

- ابتدا از تابع ابعاد کاهش یافته (PCA یا t-SNE) بر روی داده استفاده شده و داده‌ها به فضای دو بعدی منتقل شده‌اند.

- سپس از الگوریتم Agglomerative Clustering بر روی داده‌های کاهش یافته، با تعداد خوشه‌های مشخص شده (در اینجا n_clusters=2) استفاده شده است.

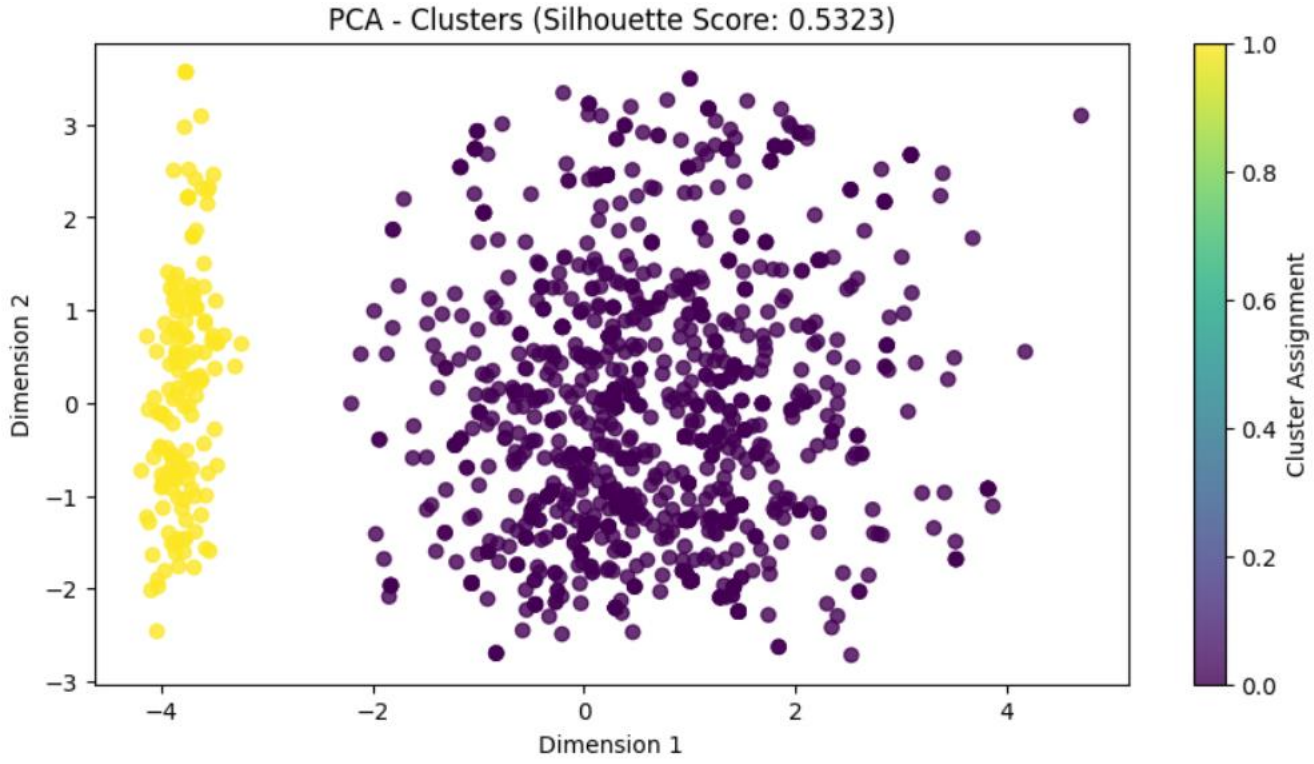
- نمودار نتیجه این خوشه‌بندی بر روی فضای دو بعدی به همراه مقدار Silhouette Score برای ارزیابی کیفیت خوشه‌بندی به تصویر کشیده شده است.

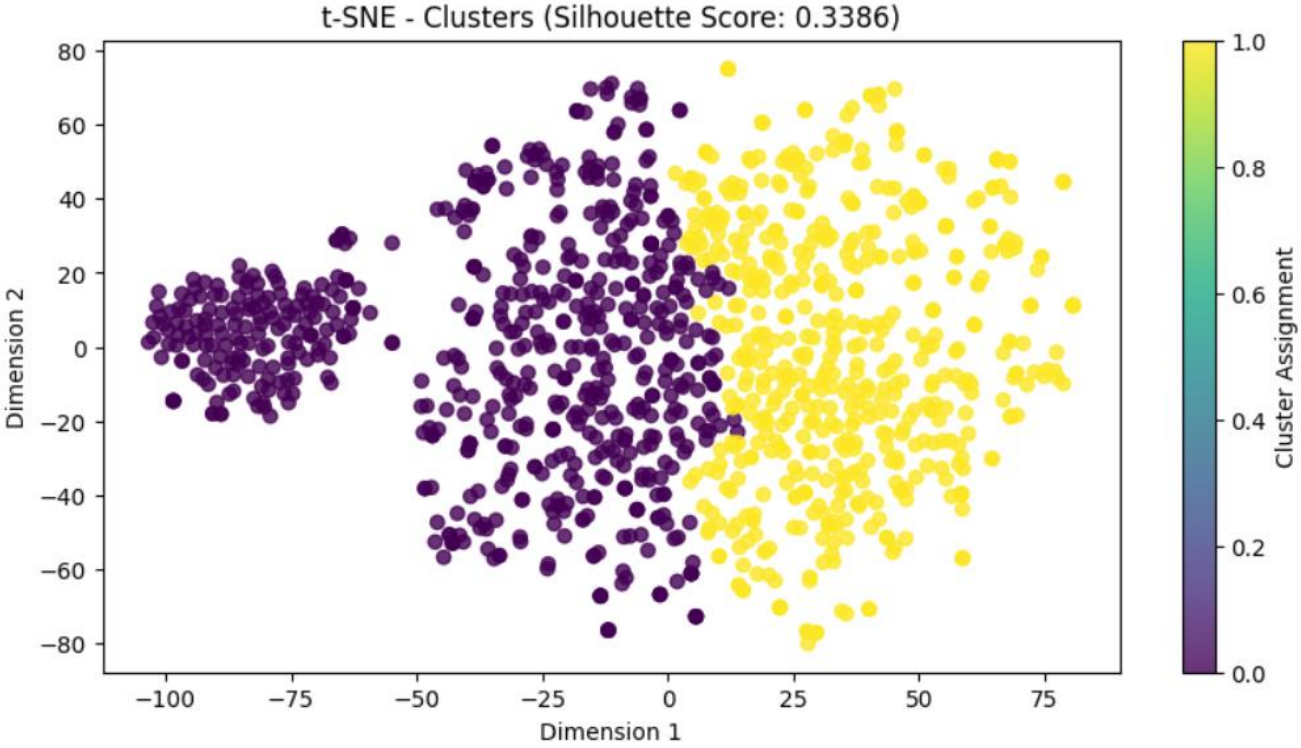
در هر نمودار:

- هر نقطه نمایانگر یک نمونه از داده است.

- رنگ هر نقطه نشان‌دهنده خوشه‌ای است کـه الگـوریتم Agglomerative Clustering آن را بـه آن اختصاص داده است.

- مقدار Silhouette Score در عنوان نمودار ذکر شده است که نشان‌دهنده کیفیت خوشه‌بندی اسـت. مقدار بالاتر از ۰ نشان‌دهنده تفکیک بهتر خوشه‌ها است.





DB-Scan -۳-۴-۲

این کد نیز دو تابع به نام‌های PCA و t-SNE را برای کاهش ابعاد داده اجرا کرده و نتایج را با استفاده از الگوریتم DBSCAN برای خوشه‌بندی کرده است. سپس نمودارهای دو بعدی از داده‌ها با توجه به نتایج خوشه‌بندی رسم شده‌اند.

برای هر تابع (PCA یا t-SNE):

- ابتدا از تابع ابعاد کاهش یافته (PCA یا t-SNE) بر روی داده‌ها استفاده شده و داده‌ها به فضای دو بعدی منتقل شده‌اند.

- سپس از الگوریتم DBSCAN بر روی داده‌های کاهش یافته، با استفاده از پارامترهای `eps` و `min_samples` مشخص شده، خوشه‌بندی انجام شده است.

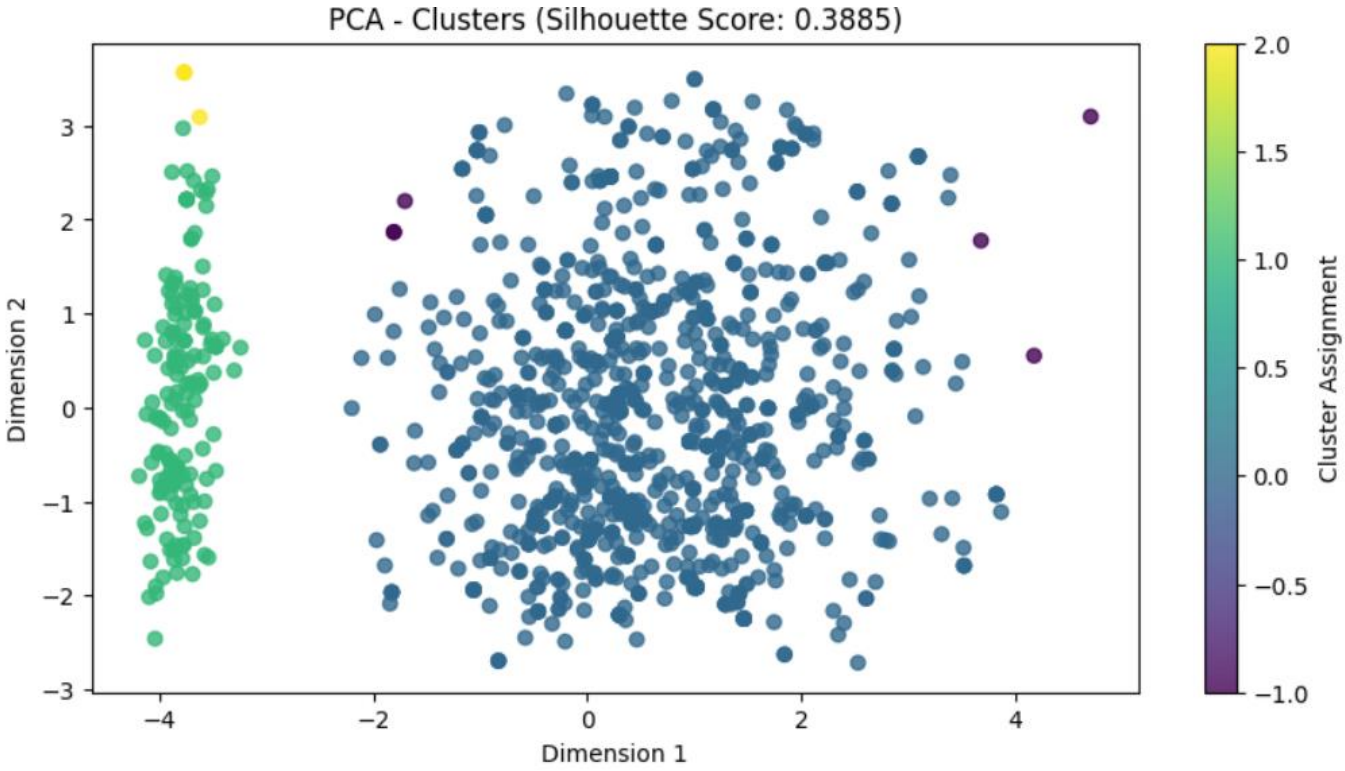
- نمودار نتیجه این خوشه‌بندی بر روی فضای دو بعدی به همراه مقدار Silhouette Score برای ارزیابی کیفیت خوشه‌بندی به تصویر کشیده شده است.

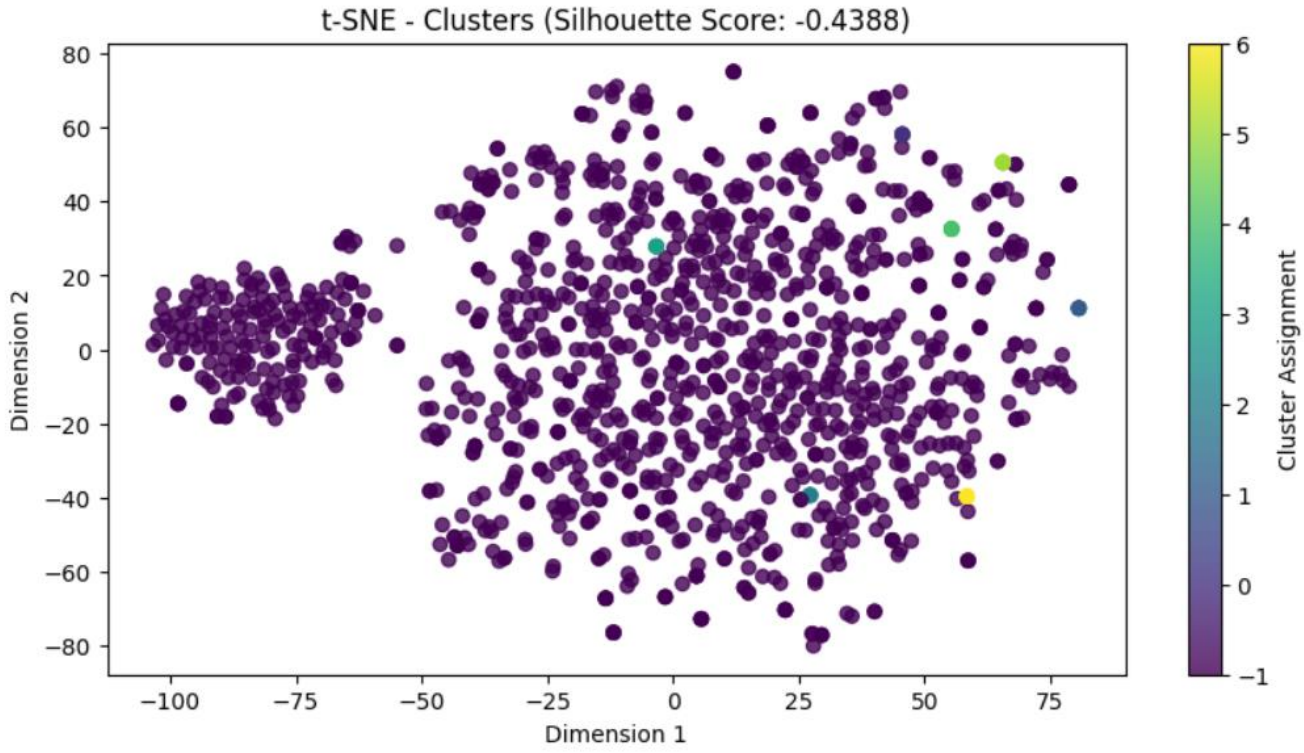
در هر نمودار:

- هر نقطه نمایانگر یک نمونه از داده است.

- رنگ هر نقطه نشان‌دهنده خوشه‌ای است که الگوریتم DBSCAN آن را به آن اختصاص داده است.

- مقدار Silhouette Score در عنوان نمودار ذکر شده است که نشان‌دهنده کیفیت خوشه‌بندی است. مقدار بالاتر از ۰ نشان‌دهنده تفکیک بهتر خوشه‌ها است.




t-SNE - Clusters (Silhouette Score: -0.4388)
Dimension 2
Dimension 1
Cluster Assignment
80
60
40
20
0
−20
−40
−60
−80
−100
−75
−50
−25
0
25
50
75
6
5
4
3
2
1
0
−1

فصل سوم

جمع‌بندی

مقایسه عملکرد

در این بخش ، معیارهای مختلف ارزیابی عملکرد الگوریتم‌های خوشه‌بندی بـرای مـدل‌های K-Means، Agglomerative Clustering و DBSCAN محاسبه شده‌اند. این معیارها به شرح زیر هستند:

۱. Silhouette Score:

- بـرای هـر الگـوریتم (K-Means، Agglomerative، DBSCAN) مقـدار Silhouette Score محاسبه شده است. این معیار نشان‌دهنده تفکیک داده‌ها به خوشه‌ها و مجزایی این خوشه‌ها است.

۲. Calinski-Harabasz Index:

- برای هر الگوریتم (K-Means، Agglomerative، DBSCAN) این معیار محاسبه شده است. این معیار به عنوان یک اندازه‌گیری برای کیفیت تفکیک خوشه‌ها و انحراف مجزایـی آن‌هـا از یکـدیگر عمـل می‌کند.

۳. Davies-Bouldin Index:

- برای هر الگوریتم (K-Means، Agglomerative، DBSCAN) این معیار محاسبه شده است. این معیار نیز به عنوان یک اندازه‌گیری از کیفیت تفکیک خوشه‌ها و انحراف مجزایی آن‌ها از یکدیگر استفاده می‌شود.

۴. Adjusted Rand Index:

- برای هر الگوریتم (K-Means، Agglomerative، DBSCAN) این معیار محاسبه شده است. این معیار برای اندازه‌گیری تطابق برچسب‌های تخمین زده شده توسط الگوریتم خوشه‌بندی با برچسـب‌های واقعی داده‌ها استفاده می‌شود.

نتایج این معیارها برای هر الگوریتم در یک جدول مقایسه گردآوری شده و با استفاده از کتابخانه Pandas به صورت جدولی چاپ شده‌اند. این جدول مقایسه این اطلاعات را برای تمام الگوریتم‌ها در یک نمای کلی نمایش می‌دهد.

Comparison Table:

	Metric	K-Means	Agglomerative	DBSCAN
0	Silhouette Score	0.076649	0.165082	-0.045398
1	Calinski-Harabasz Index	59.211600	1.939724	6.669355
2	Davies-Bouldin Index	2.891287	0.711163	1.043942
3	Adjusted Rand Index	0.018327	0.000000	0.066177

منابع و مراجع

[1] https://www.kaggle.com/datasets/heeraldedhia/groceries-dataset

پیوست‌ها

- **https://colab.research.google.com/drive/1uTHqSQkcUfYTR_iR33yGOBf6JrR6SldU?usp=sharing**

Abstract

This project performed cluster analysis on the insurance dataset. First, we divided the data into different clusters using different methods such as KMeans, Agglomerative Clustering, and DBSCAN. Then we compared the clustering results using evaluation criteria such as Silhouette Score. Also, the sensitivity of the model to changes in parameters was investigated. Next, dimensionality reduction methods such as PCA and t-SNE were used to visualize the data. These visualizations show the strengths and weaknesses of each algorithm along with comparing the results using quality metrics. Finally, these results are used to analyze and present a comparative table of the performance of each clustering algorithm in the article.

Key Words:

Clustering, fraud detection, K-means, DB-Scan, Agglomerative Clustering

Amirkabir University of Technology
(Tehran Polytechnic)

Department of Computer Science

Project 6

Advanced Methods in Clustering

By
Samin Mahdipour

Supervisor
Dr.Ghatee

Advisor
Dr.Yousofi Mehr

November 2023