

داكيمونت پروژه اول مبانی هوش محاسباتی

سید حمیدرضا حسینی

محمد اشکوه

1403-1404

فاز ۱ : Feature Extraction

1.1 مقدمه

در این مرحله، ویژگی‌های مختلفی از تصاویر استخراج شده‌اند تا در مراحل بعدی، از آن‌ها برای خوشبندی تصاویر استفاده شود. ویژگی‌های انتخاب شده شامل اطلاعات رنگی، وضوح تصویر، چگالی لبه‌ها، ویژگی‌های آماری و ویژگی‌های بافتی هستند.

1.2 ویژگی‌ها

HSV ویژگی رنگی

ویژگی‌های رنگی HSV یکی از ویژگی‌های پرکاربرد در پردازش تصویر است که با استفاده از تبدیل تصویر به فضای رنگی Hue، Saturation، Value استخراج می‌شود. این ویژگی‌ها می‌توانند اطلاعات دقیق‌تری از رنگ‌ها و ویژگی‌های تصویری نسبت به فضای رنگی RGB فراهم کنند. این ویژگی‌ها شامل هیستوگرام رنگ‌ها در سه کanal (رنگ)، (اشیاع رنگ) و (درخشندگی) هستند و می‌توانند برای شناسایی انواع مختلف ویژگی‌های تصویری مانند رنگ‌ها و شدت آن‌ها به کار روند.

مراحل محاسبه:

1. تبدیل تصویر از فضای رنگی BGR به HSV.
2. تقسیم تصویر به سه کanal HSV
 - (H) برای نمایش رنگ.
 - (S) برای نمایش شدت رنگ.
 - (V) برای نمایش روشنایی.
3. محاسبه هیستوگرام هر کدام از کanal‌ها با تعداد مشخصی از بین‌ها (`hist_bins`) که نشان‌دهنده توزیع رنگ‌ها در هر کanal است.
4. نرمال‌سازی هیستوگرام‌ها برای مقیاس‌دهی و کاهش اثر مقادیر نوردهی یا اندازه تصویر.

5. محاسبه میانگین و انحراف معیار برای هر کanal برای داشتن درک بیشتری از شدت رنگ و روشنابی.

♦ تفسیر مقادیر:

• کanal H : (Hue)

- نشان‌دهنده رنگ تصویر است. مقادیر مختلف Hue به رنگ‌های مختلف مانند قرمز، آبی، سبز و ... اشاره می‌کند.
- اگر مقادیر Hue متنوع باشد، تصویر ممکن است دارای رنگ‌های مختلف و متنوعی باشد.

• کanal S : (Saturation)

- نشان‌دهنده شدت یا غنی بودن رنگ است. مقادیر بالاتر Saturation به رنگ‌های پررنگ و اشباع‌شده اشاره دارند.
- اگر مقادیر Saturation کم باشد، تصویر رنگ‌های کمرنگ یا خنثی خواهد داشت.

• کanal V : (Value)

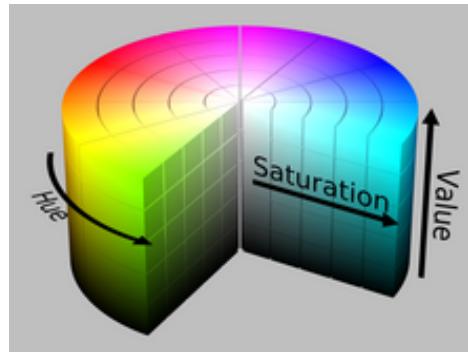
- نشان‌دهنده روشنابی تصویر است. مقادیر بالاتر Value به تصویر روشن‌تر و مقادیر پایین‌تر به تصویر تاریک‌تر اشاره دارند.

♦ دلیل انتخاب:

ویژگی‌های HSV به دلیل اینکه از فضای رنگی متقاوی نسبت به فضای RGB استفاده می‌کند، می‌توانند تشخیص رنگ‌ها و ویژگی‌های تصویری را با دقت بیشتری انجام دهند. کanal‌های Value، Hue و Saturation به طور جداگانه قادر به ارائه اطلاعات مهم درباره رنگ، شدت و روشنابی تصویر هستند. این ویژگی‌ها معمولاً در شناسایی اشیاء، تشخیص الگوهای رنگی، و تقسیم تصاویر با رنگ‌های متنوع به کار می‌روند.

♦ مثال:

تصویری از یک گل با رنگ‌های متنوع در مقابل یک تصویر از یک سطح خاکی با رنگ‌های کمرنگ. ویژگی‌های HSV می‌توانند به تقسیم رنگ‌های مختلف گل از رنگ خاک کمک کنند.



۴. پوشح تصویر (Laplacian Variance)

یک ویژگی عددی در تحلیل تصاویر است که با استفاده از فیلتر لابلسین (Laplacian Filter) بر روی تصویر خاکستری محاسبه می‌شود و نشان‌دهنده میزان پوشح (sharpness) تصویر می‌باشد. این ویژگی به عنوان شاخصی برای تشخیص میزان تمرکز (focus) یا تاری (blur) تصویر کاربرد دارد.

مراحل محاسبه:

1. تبدیل تصویر به فضای خاکستری (Grayscale)
2. اعمال فیلتر لابلسین (Laplacian Operator)
3. اعمال فیلتر لابلسین (Laplacian Operator)

تفسیر مقادیر:

- اگر مقدار Laplacian Variance زیاد باشد: تصویر دارای لبه‌های قوی و تغییرات زیاد روشناختی است → تصویر واضح (sharp) است.
- اگر مقدار Laplacian Variance کم باشد: تصویر فقد تغییرات شدید روشناختی و فقد لبه‌های مشخص است → تصویر تار (blurry) است.
- دلیل انتخاب: این ویژگی به شناسایی میزان پوشح تصویر کمک می‌کند. تصاویری که دارای جزئیات زیاد یا لبه‌های واضح هستند، واریانس بالاتری دارند. بنابراین، این ویژگی می‌تواند تصاویر واضح و دقیق را از تصاویر تار و مبهم متمایز کند.

مثال: تصویری از یک منظره طبیعی واضح با جزئیات فراوان در مقابل یک تصویر تار یا کمپوش.

چگالی لبه‌ها (Edge Density) 🔍

Edge Density یا «تراکم لبه‌ها»، یکی از ویژگی‌های ساختاری (structural features) در تحلیل تصویر است که بیان می‌کند چه مقدار از تصویر دارای لبه‌های شناسایی شده است. این ویژگی نشان‌دهنده‌ی پیچیدگی بصری یا میزان جزئیات ساختاری در تصویر می‌باشد.

نحوه محاسبه:

1. تبدیل تصویر به خاکستری (Grayscale)
2. استخراج لبه‌ها با الگوریتم Canny
3. محاسبه تراکم لبه‌ها (Edge Density)

تفسیر مقادیر

- **مقدار پایین (۰.۰۰۰ تا ۰.۱۵):** این مقدار نشان‌دهنده‌ی تصاویری است که لبه‌های کمی دارند و بهطور معمول شامل تصاویر ساده یا تار هستند. به عنوان مثال، تصویری از یک آسمان صاف یا یک دیوار یکنواخت که هیچ تغییرات شدیدی در شدت روشنایی ندارد، Edge Density پایینی خواهد داشت.
- **مقدار متوسط (۰.۱۵ تا ۰.۴۰):** در این محدوده، تصویر شامل لبه‌های کمی است که جزئیات محدود دارند. مثلاً یک منظره طبیعی ساده با جزئیات کم، مانند یک میدان یا یک فضای باز با لبه‌های غیر واضح.
- **مقدار بالا (۰.۴۰ تا ۰.۶۵):** تصویری با این مقدار Edge Density، جزئیات و لبه‌های بیشتری دارند. به عنوان مثال، یک تصویر از یک منظره شهری یا یک تصویر پزشکی که دارای لبه‌های دقیق‌تر و مشخص‌تر است.
- **مقدار بسیار بالا (۰.۶۵ تا ۱.۰۰):** در این حالت، تصویر دارای لبه‌های بسیاری است و ممکن است پیچیدگی زیادی داشته باشد. این تصاویر معمولاً جزئیات زیاد و تغییرات زیادی در شدت روشنایی دارند، مانند تصاویری از یک شهر با ساختمان‌های متراکم یا تصاویری که ساختار پیچیده‌تری دارند.

♦ **دلیل انتخاب:** این ویژگی کمک می‌کند تا تصاویری با جزئیات زیاد از تصاویر ساده متمایز شوند.

📍 **مثال:** تصویری از یک ساختمان پیچیده در مقابل یک صفحه سفید.

(ویژگی‌های آماری) Statistical Features 📈

ویژگی‌های آماری شامل توزیع شدت روشنایی در تصویر هستند که می‌توانند اطلاعاتی درباره توزیع نور، کنتراست، و یکنواختی تصویر به دست دهند. این ویژگی‌ها برای شناسایی تصاویری با توزیع یکنواخت روشنایی یا تغییرات زیاد در روشنایی استفاده می‌شوند.

مراحل محاسبه:

1. تبدیل تصویر به فضای خاکستری (Grayscale)
2. محاسبه میانگین و واریانس شدت روشنایی

تفسیر مقادیر:

- اگر مقدار میانگین روشنایی بالا باشد: تصویر روشن‌تر است.
- اگر مقدار واریانس زیاد باشد: تصویر دارای تقاؤت‌های روشنایی و کنتراست زیادی است.
- اگر مقدار میانگین و واریانس کم باشد: تصویر یکنواخت و کمکنتراست است.

♦ دلیل انتخاب: این ویژگی‌ها به تشخیص تصاویر با توزیع روشنایی یکنواخت یا تصاویری که تغییرات روشنایی و کنتراست زیادی دارند کمک می‌کند.

♦ مثال: تصویری از یک دیوار سفید یکنواخت در مقابل یک صحنه با نورپردازی پیچیده.

Color Variance (واریانس رنگ)

ویژگی‌ای است که تغییرات رنگی در تصویر را اندازه‌گیری می‌کند. این ویژگی کمک می‌کند تا تصاویری با تنوع رنگی زیاد را از تصاویری با رنگ‌های یکنواخت تمایز دهیم

مراحل محاسبه:

1. محاسبه واریانس شدت رنگی در هر یک از کانال‌های رنگی (HSV یا RGB)
2. محاسبه واریانس کلی رنگ در تصویر

تفسیر مقادیر:

- اگر مقدار Color Variance زیاد باشد: تصویر دارای تنوع رنگی بالا است → تصویر رنگارنگ است.
- اگر مقدار Color Variance کم باشد: تصویر دارای رنگهای یکنواخت است → تصویر ساده و یکنواخت است.

♦ دلیل انتخاب: این ویژگی به شناسایی تصاویری که تنوع رنگی دارند کمک می‌کند، که می‌تواند در شناسایی تصاویر طبیعی یا تصاویر با بافت‌های رنگی پیچیده مفید باشد.

مثال: تصویری از یک گلزار رنگارنگ در مقابل یک تصویر از یک صفحه خاکی یکنواخت. 

(زبری) Roughness

ویژگی‌ای است که تغییرات ناگهانی در شدت روشنایی تصویر را اندازمگیری می‌کند. این ویژگی برای شناسایی سطوح ناصاف و زبر در تصویر مفید است.

♦ مراحل محاسبه:

1. تبدیل تصویر به فضای خاکستری (Grayscale)
2. محاسبه تقاضاهای شدت روشنایی در پیکسل‌های مجاور
3. محاسبه انحراف معیار تقاضاهای به عنوان زبری

♦ تفسیر مقادیر:

- اگر مقدار Roughness زیاد باشد: تصویر دارای سطوح ناصاف و تغییرات زیادی است → تصویر زبر و ناصاف است.
- اگر مقدار Roughness کم باشد: تصویر دارای سطوح صاف و یکنواخت است → تصویر نرم و صاف است.

♦ دلیل انتخاب: این ویژگی می‌تواند در شناسایی سطوح با بافت‌های پیچیده و ناصاف از سطوح صاف و یکنواخت کاربرد داشته باشد.

مثال: تصویری از یک سطح صخره‌ای یا دیوار آجری در مقابل یک سطح صاف و صیقلی. 

Entropy (آنتروپی)

ویژگی‌ای است که پیچیدگی و تصادفی بودن تصویر را اندازه‌گیری می‌کند. این ویژگی نشان‌دهنده میزان اطلاعات و جزئیات موجود در تصویر است.

مراحل محاسبه:

1. محاسبه توزیع احتمال شدت‌های روشنایی در تصویر
2. استفاده از فرمول آنتروپی برای اندازه‌گیری پیچیدگی و تصادفی بودن

تفسیر مقادیر:

- اگر مقدار Entropy زیاد باشد: تصویر دارای جزئیات پیچیده و ساختار تصادفی است → تصویر پیچیده است.
- اگر مقدار Entropy کم باشد: تصویر دارای ساختار ساده و یکنواخت است → تصویر ساده است.

♦ دلیل انتخاب: این ویژگی به تشخیص تصاویری که دارای اطلاعات زیاد و پیچیدگی بالا هستند کمک می‌کند، که در پردازش تصویر و شناسایی الگوهای پیچیده کاربرد دارد.

📍 مثال: تصویری از یک پیچیدگی بافت یا زمینه تصادفی در مقابل یک تصویر ساده و یکنواخت.

Homogeneity (همگنی)

ویژگی‌ای است که میزان یکنواختی بافت در تصویر را اندازه‌گیری می‌کند. این ویژگی برای شناسایی تصاویر با بافت‌های یکدست مفید است.

مراحل محاسبه:

1. محاسبه ماتریس هم‌همواری (GLCM) برای تصویر خاکستری
2. استخراج مقدار همگنی از ماتریس هم‌همواری

تفسیر مقادیر:

- اگر مقدار Homogeneity زیاد باشد: تصویر دارای بافت‌های یکنواخت و ساده است → تصویر ساده و یکنواخت است.

- اگر مقدار Homogeneity کم باشد: تصویر دارای تغییرات زیادی در بافت است → تصویر پیچیده است.

♦ دلیل انتخاب: این ویژگی برای شناسایی تصاویری که دارای بافت‌های یکنواخت هستند و تشخیص تصاویری با بافت‌های پیچیده کاربرد دارد.

مثال: تصویری از یک سطح صاف و شیشه‌ای در مقابل یک تصویر با تغییرات بافت پیچیده

فاز 2: Feature Selection

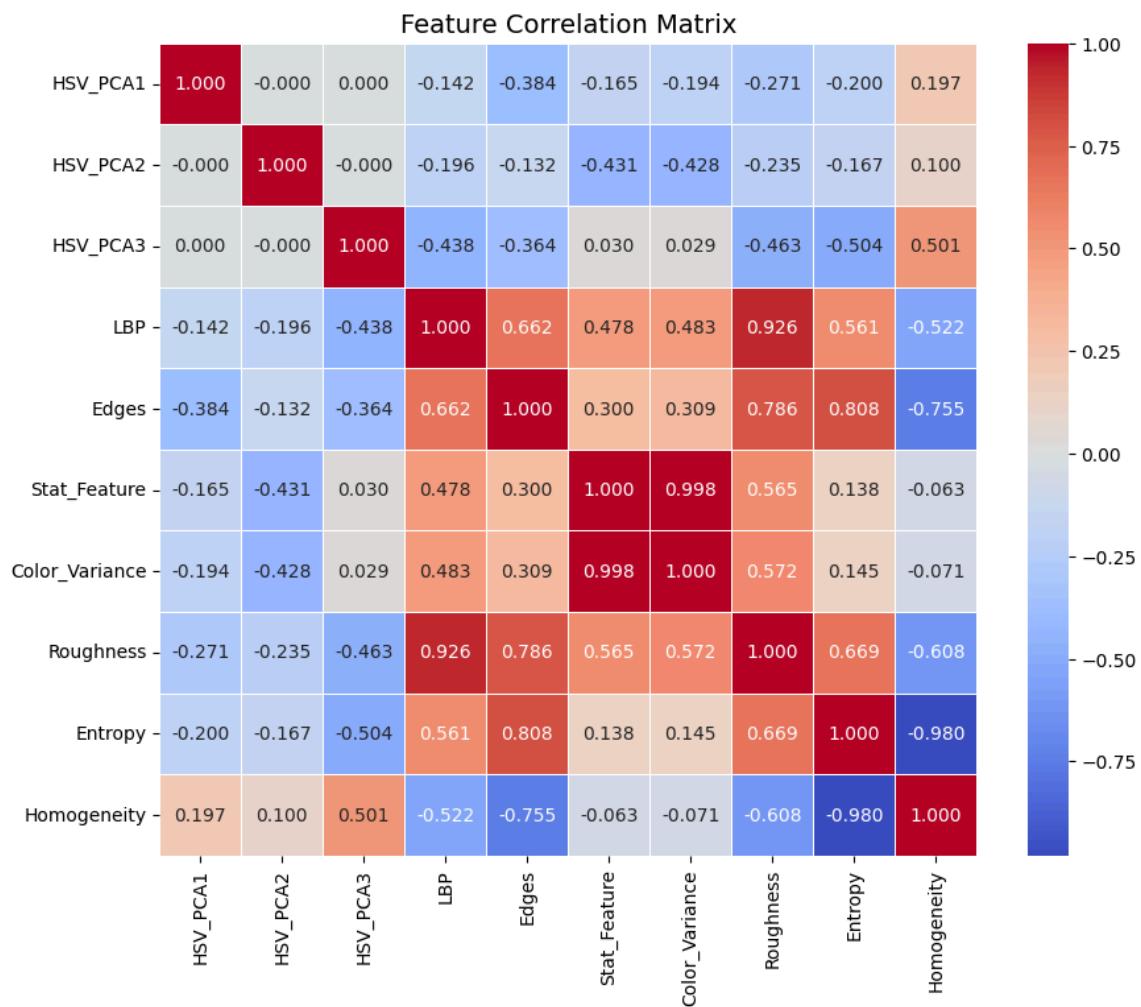
در این فاز هدف آن است که از میان ویژگی‌های استخراج شده در فاز اول، آن دسته از ویژگی‌هایی انتخاب شوند که بیشترین تأثیر را در فرآیند خوشبندی دارند و کمترین همبستگی ممکن را با یکدیگر دارند. این کار به بهبود عملکرد الگوریتم‌های خوشبندی و کاهش بعد داده‌ها کمک می‌کند.

2.1 آماده‌سازی داده

ویژگی‌هایی که در فاز اول استخراج شدند، در قالب یک فایل CSV به نام `features.csv` ذخیره شده‌اند. ابتدا این فایل را لود می‌کنیم

2.2 محاسبه ماتریس همبستگی

نتیجه‌ی این مرحله یک ماتریس مربعی بود که هر خانه‌ی آن میزان همبستگی بین دو ویژگی را نمایش می‌دهد.



2.3 حذف ویژگی‌های با همبستگی بالا

سه روش برای انتخاب ویژگی‌ها پیاده‌سازی شده است:

روش اول: Threshold-based

در این روش، ویژگی‌هایی که همبستگی بیش از مقدار آستانه با سایر ویژگی‌ها داشتند، حذف شدند. هدف حذف ویژگی‌هایی بود که اطلاعات مشابه منتقل می‌کنند. پس از این مرحله، تنها ویژگی‌هایی باقی ماندند که با یکدیگر همبستگی پایینی داشتند.

خروجی:

Best k: 4 | silhouette score: 0.38 | Features: ['HSV_PCA1', 'HSV_PCA2', 'HSV_PCA3', 'Homogeneity']

ویژگی‌های انتخاب شده و دلیل انتخاب آن‌ها

HSV_PCA1 .1

● چیست؟

اولین مؤلفه اصلی از کاهش بعد (PCA) بر روی هیستوگرام رنگی HSV تصویر.

● چرا انتخاب شده؟

این مؤلفه بیشترین و مهمترین اطلاعات رنگی تصویر را به صورت فشرده در خود دارد.

● چه کمکی می‌کند؟

باعث می‌شود تصاویر با رنگ غالب مشابه (مثلًا آبی برای دریا یا سبز برای جنگل) در خوش‌های مشابه قرار بگیرند.

HSV_PCA2 .2

● چیست؟

دومین مؤلفه اصلی پس از کاهش بعد داده‌های رنگی HSV.

● چرا انتخاب شده؟

اطلاعات رنگی مکمل HSV_PCA1 را شامل می‌شود و تقاضاهای رنگی درون تصاویر با رنگ غالب یکسان را مشخص می‌کند.

● چه کمکی می‌کند؟

در تفکیک تصاویر با رنگ‌های نزدیک ولی ترکیب متقاوت مؤثر است؛ مثلاً تمایز بین آسمان آبی و دریا.

HSV_PCA3 .3

● چیست؟

سومین مؤلفه حاصل از PCA روی ویژگی‌های رنگی.

● چرا انتخاب شده؟

شامل جزئیات دقیق‌تر رنگ، مانند شدت نور، سایه‌ها یا تضادهای رنگی داخل تصویر است.

● چه کمکی می‌کند؟

در تفکیک تصاویر پیچیده‌تر (مثلاً جنگل با نقاط نورانی یا سایه‌دار) نقش تکمیلی دارد.

Homogeneity .4

● چیست؟

ویژگی بافتی به‌دست‌آمده از ماتریس هم‌زنایی خاکستری (GLCM)، که میزان یکنواختی تصویر را اندازه‌گیری می‌کند.

● چرا انتخاب شده؟

باft، مکمل رنگ در تحلیل تصاویر طبیعت است و این ویژگی تمایز واضحی بین مناطق صاف و بافتدار ایجاد می‌کند.

● چه کمکی می‌کند؟

به مدل کمک می‌کند تا مثلاً تقاضه بین بیابان (بافت یکنواخت) و جنگل (بافت پیچیده) را تشخیص دهد، حتی اگر رنگ‌ها مشابه باشند.

جمع‌بندی:

- این چهار ویژگی انتخاب شده‌اند زیرا با هم:
- ترکیبی کامل از رنگ و بافت ارائه می‌دهند،
 - کمترین همبستگی را با هم دارند (طبق محاسبه ماتریس همبستگی)،
 - و در کنار هم اطلاعات کافی برای تفکیک دقیق تصاویر طبیعی فراهم می‌کنند.

روش دوم: نگداشتن کم همبسترنین ویژگی‌ها

در این روش، هدف نهایی انتخاب K ویژگی‌ای است که کمترین همبستگی ممکن را با یکدیگر دارند. مراحل کلی این الگوریتم به شکل زیر انجام می‌شود:

- محاسبه‌ی همه‌ی همبستگی‌های ممکن بین جفت ویژگی‌ها و ذخیره آن‌ها در یک لیست.
- مرتب‌سازی لیست بر اساس بیشترین مقدار همبستگی
- در هر مرحله، بررسی اینکه آیا هر دو ویژگی دارای همبستگی بالا هستند، اگر بله، یکی از آن‌ها حذف می‌شود.
- این فرآیند تا زمانی ادامه پیدا می‌کند که تنها K ویژگی باقی بماند.

این الگوریتم به نوعی حریصانه (**Greedy**) عمل می‌کند، به این صورت که در هر مرحله پر همبستگی‌ترین جفت‌ها را پیدا کرده و یکی را حذف می‌کند تا به کمترین همبستگی بین باقی‌مانده‌ها برسد.

نتیجه: این روش از روش اول و سوم کمتر بود.

روش سوم: انتخاب بهترین زیرمجموعه با استفاده از **Silhouette Score** (بررسی همه زیرمجموعه و انتخاب بهترین)

در این روش ترکیب‌های مختلفی از زیرمجموعه‌ها (حداقل ۲ و حداقل تمام ویژگی‌ها) بررسی شدند. برای هر ترکیب، الگوریتم KMeans با تعداد خوش‌های مختلف اجرا شد و سیلوئت اسکور برای آن محاسبه شد. ضمناً این الگوریتم فقط حالت‌هایی را بررسی می‌کرد که بهترین k در آنها بین ۴ تا ۷ باشد (به عنوان مثال اگر غیر این بود مجموعه‌هایی با score: 58 بدست می‌آمد ولی بهترین k در آنها برابر ۲ بود).

خروجی:

Best k: 5 | silhouette score: 0.42 | Features: ('HSV_PCA2', 'HSV_PCA3', 'Edges')

با توجه به هزینه زمانی بالا، این روش برای بررسی جامع‌تر انتخاب نشده.

در کل متد اول به عنوان متد نهایی انتخاب شد

2.4 ذخیره‌سازی ویژگی‌های منتخب

ویژگی‌های نهایی انتخاب‌شده در فاز دوم، در فایل `features_selected.csv` ذخیره شدند و مبنای فاز‌های بعدی (خوبه‌بندی، ارزیابی و پیش‌بینی) قرار گرفتند.

فاز ۳ و ۴ : Clustering & Visualization

3.1 پیش‌پردازش داده‌ها

مقیاس‌بندی (Normalization)

جهت اطمینان از اینکه تمامی ویژگی‌ها در یک مقیاس یکسان قرار دارند و تقاضات‌های مقیاسی تأثیر منفی بر روی الگوریتم‌های خوبه‌بندی نداشته باشد، داده‌ها با استفاده از استانداردسازی (StandardScaler) نرمال می‌شوند. این مرحله باعث می‌شود که الگوریتم‌هایی مانند K-Means بر اساس فاصله‌های اقلیدسی دقیق‌تر عمل کنند.

3.2 اجرای الگوریتم‌های خوشبندی

برای ارزیابی عملکرد خوشبندی، از سه الگوریتم مختلف استفاده شده است:

K-Means 3.2.1

رویه اجرای الگوریتم:

الگوریتم K-Means داده‌های نرمال‌شده را به k خوشه تقسیم می‌کند. ابتدا با بررسی چند مقدار به کمک دو روش Elbow (محاسبه SSE) و محاسبه Silhouette Score، تعداد بهینه‌ی خوشه مشخص می‌شود. ابتدا به معرفی دو متغیر داریم.

Elbow 3.2.1.1 روش

یک تکنیک بصری برای تعیین تعداد بهینه‌ی خوشه‌ها در الگوریتم‌هایی مانند K-Means است. هدف این روش، یافتن نقطه‌ای است که پس از آن، افزایش تعداد خوشه‌ها تأثیر قابل توجهی در بهبود خوشبندی ندارد.

نحوه عملکرد:

در این روش، برای مقادیر مختلف k (تعداد خوشه‌ها)، مقدار خطای خوشبندی یا مجموع مربعات خط (**SSE - Sum of Squared Errors**) محاسبه می‌شود. سپس این مقادیر در قالب یک نمودار نمایش داده می‌شوند.

در محور افقی تعداد خوشه‌ها (k) و در محور عمودی SSE قرار می‌گیرد. با افزایش k ، مقدار SSE کاهش می‌یابد، اما از یک نقطه به بعد، کاهش SSE کند می‌شود. این نقطه، به دلیل شکل خمیدگی نمودار، به نام elbow شناخته می‌شود و تعداد خوشه‌ی مناسب را نشان می‌دهد.

کاربرد:

• تعیین تعداد بهینه‌ی خوشه‌ها برای الگوریتم K-Means.

• جلوگیری از تقسیم بیش از حد یا کم خوشبندی داده‌ها.

3.2.1.2 امتیاز Silhouette (سیلوئت)

تعريف:

امتیاز سیلوئت معیاری عددی برای سنجش کیفیت خوشنده‌بندی است که نحوه قرارگیری هر نمونه را نسبت به خوشنده خودش و سایر خوشنده‌ها بررسی می‌کند.

نحوه عملکرد:

برای هر نمونه در داده:

- **a**: میانگین فاصله‌ی آن نمونه تا سایر نقاط در همان خوشه.
- **b**: میانگین فاصله‌ی آن نمونه تا نزدیکترین خوشه‌ی دیگر (یعنی خوشه‌ای که به آن تعلق ندارد).

امتیاز سیلوئت از رابطه زیر به دست می‌آید:

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

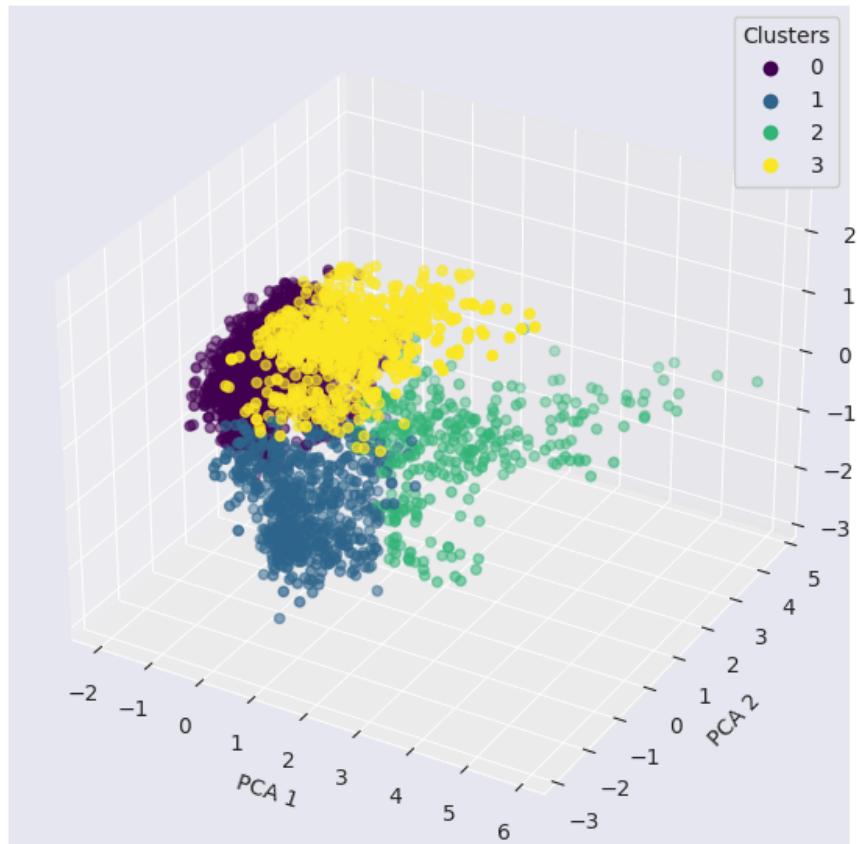
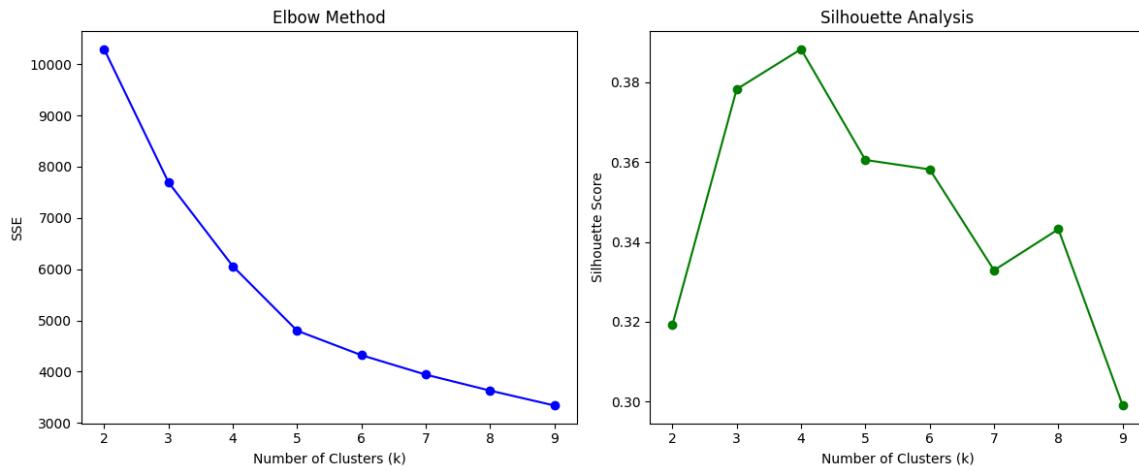
این امتیاز بین -1 تا 1 قرار دارد:

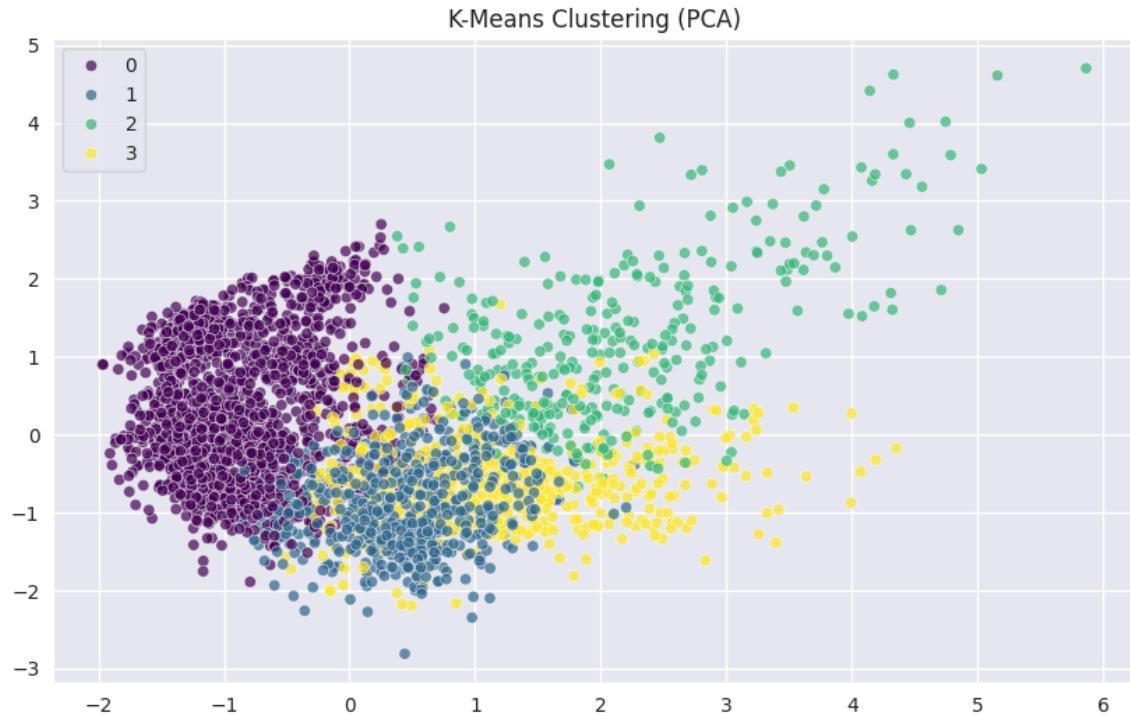
- **نزدیک به 1**: خوشنده‌بندی بسیار خوب؛ نمونه به خوشه‌ی خودش نزدیک و از سایر خوشنده‌ها دور است.
- **نزدیک به 0**: نمونه در مرز بین دو خوشه قرار دارد.
- **منفی**: نمونه احتمالاً در خوشه‌ی اشتباهی قرار گرفته است.

کاربرد:

- ارزیابی کیفیت خوشنده‌بندی بدون نیاز به برچسب‌های واقعی.
- مقایسه‌ی عملکرد الگوریتم‌های مختلف یا تعداد مختلف خوشنده‌ها.

k means نتائج





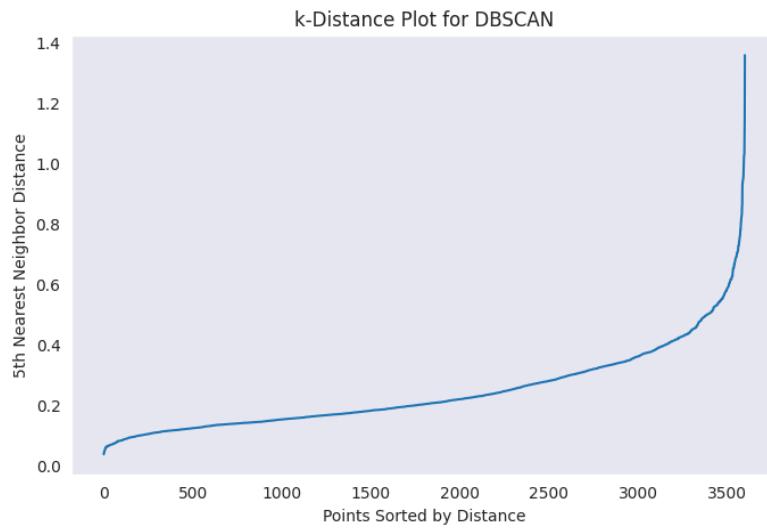
DBSCAN 3.2.2

روش کار:

DBSCAN یک الگوریتم خوشه‌بندی مبتنی بر تراکم است. در این الگوریتم، پارامترهای `eps` (فاصله) و `min_samples` (تعداد نمونه‌های مورد نیاز برای تشخیص یک خوشه) تعیین می‌شوند.

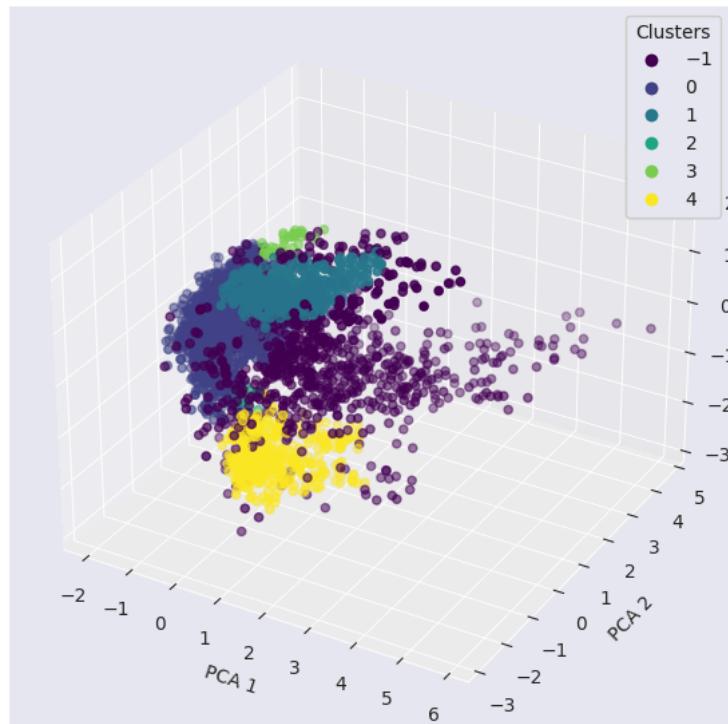
انتخاب پارامترها:

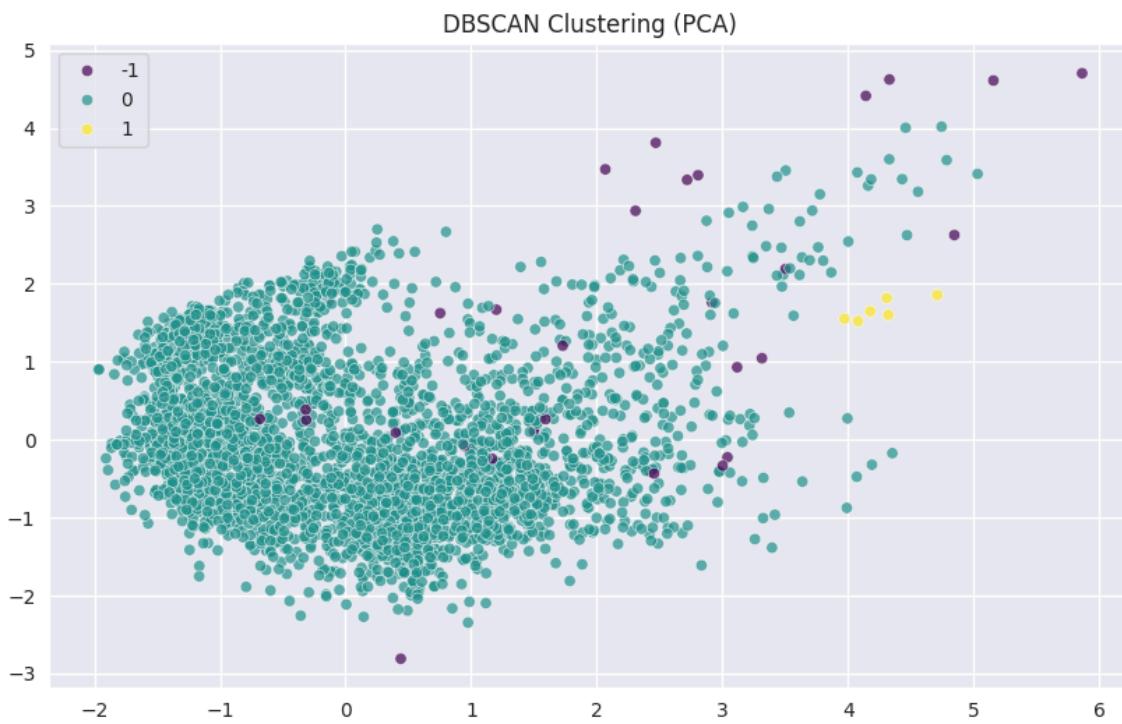
برای به دست آوردن مقدار مناسب `eps`، از نمودار `k-distance` (با استفاده از `NearestNeighbors`) استفاده می‌شود. نقاطی که بر چسب آن‌ها ۱-تشخیص داده می‌شوند به عنوان نویز در نظر گرفته شده و می‌توانند در ارزیابی نهایی جداگانه پردازش شوند.



نمودار بالا فاصله‌ی هر نقطه تا سومین نزدیکترین همسایه‌اش را نمایش می‌دهد. نقاط بر اساس این فاصله‌ها مرتب شده‌اند. با توجه به شکل نمودار، یک تغییر شیب مشخص (نقطه زانو) در حدود مقدار **0.5** دیده می‌شود. این نقطه نشان‌دهنده‌ی مرزی میان نقاط متراکم (خوش‌ها) و نقاط دورافتاده (نویز) است.

با توجه به این تحلیل، مقدار مناسب برای پارامتر **ϵ (epsilon)** در الگوریتم DBSCAN حدود **0.5** می‌باشد. این مقدار به تفکیک بهتر خوش‌ها و حذف نویز از داده‌ها کمک می‌کند.





برای خوشه‌بندی داده‌ها، ابتدا از نمودار k -Distance با $k=5$ استفاده شد تا مقدار مناسب ϵ تعیین گردد. بر اساس نقطه‌ی زانو در نمودار، مقدار $\epsilon \approx 0.75$ انتخاب شد. سپس الگوریتم DBSCAN با پارامترهای زیر اعمال گردید:

$$\epsilon = 0.5$$

$$\text{min_samples} = 3$$

نتیجه‌ی خوشه‌بندی (پس از کاهش ابعاد با PCA برای نمایش دو بعدی) به صورت زیر است:

- DBSCAN توانست فقط 6 را در داده‌ها شناسایی کند.

- برخی نقاط به عنوان نویز (برچسب -1) شناسایی شده‌اند که در نواحی با تراکم پایین قرار داشتند..

Agglomerative Clustering 3.2.3

- روش سلسله‌مراتبی:

خوشه‌بندی سلسله‌مراتبی (Agglomerative) به صورت تکراری نزدیکترین نقاط یا خوشه‌ها را به یکدیگر ادغام می‌کند تا در نهایت ساختار درختی (dendrogram) شکل بگیرد.

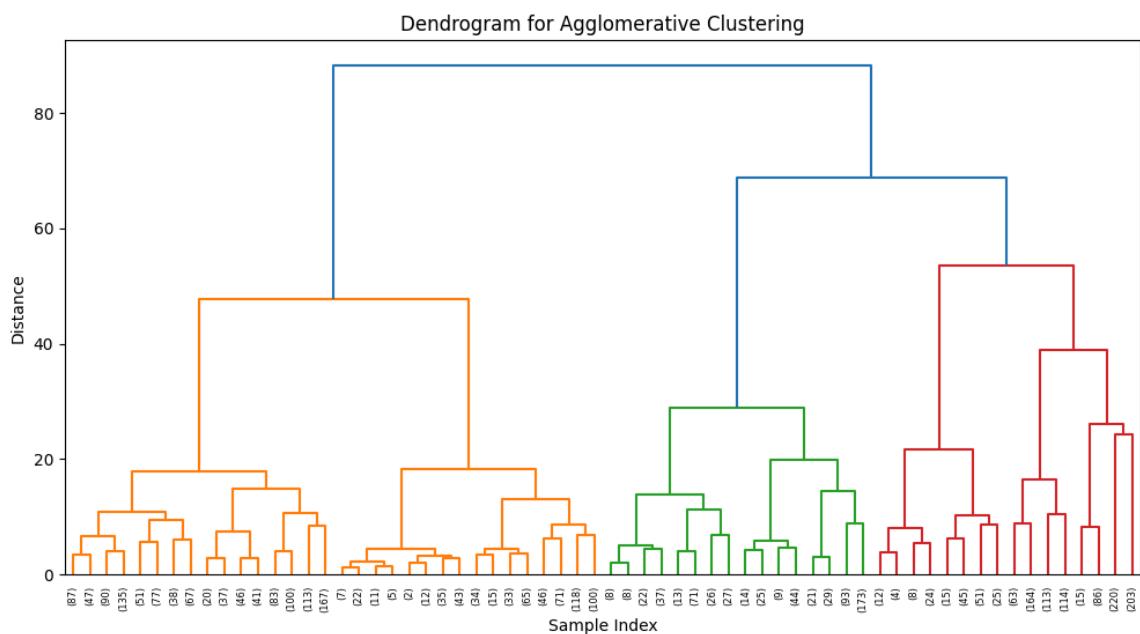
● دندروگرام:

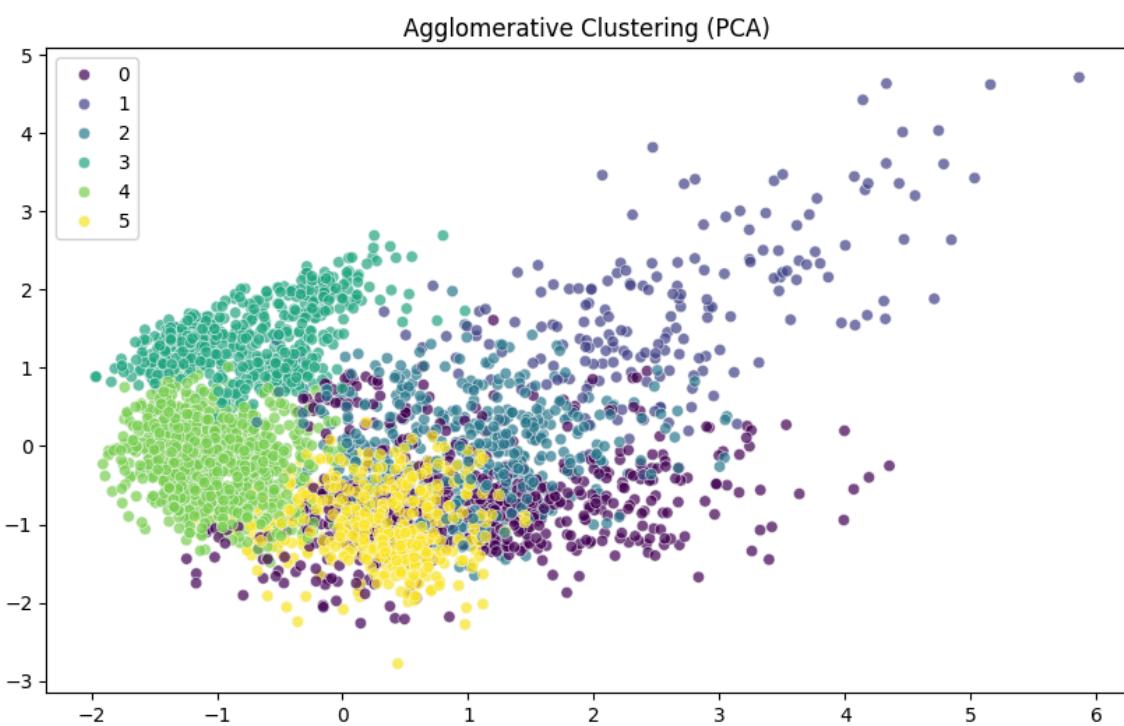
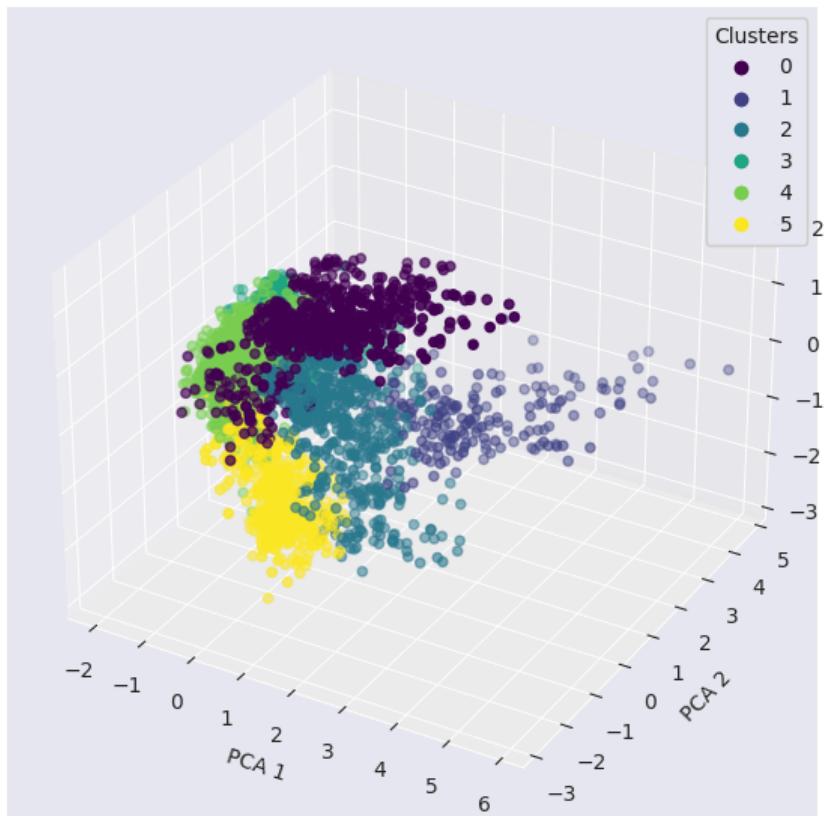
دندروگرام رسم شده نقش مهمی در بصری سازی ساختار سلسله مراتبی داده دارد و این امکان را می‌دهد تعداد خوشی مناسب را با تکیه بر فاصله‌های ادغام مشاهده کند.

3.2.3.1 دندروگرام

- دندروگرام رابطه‌ی بین نمونه‌ها را بر اساس شباهت نمایش می‌دهد.
- محور عمودی (Distance) نشان‌دهنده فاصله یا میزان ناهماهنگی بین خوشی‌های ادغام شده است.
- هر بار که دو خوشی به هم وصل می‌شوند، این اتصال در ارتقای از محور عمودی رسم می‌شود که برابر با فاصله آن‌هاست.

خروج



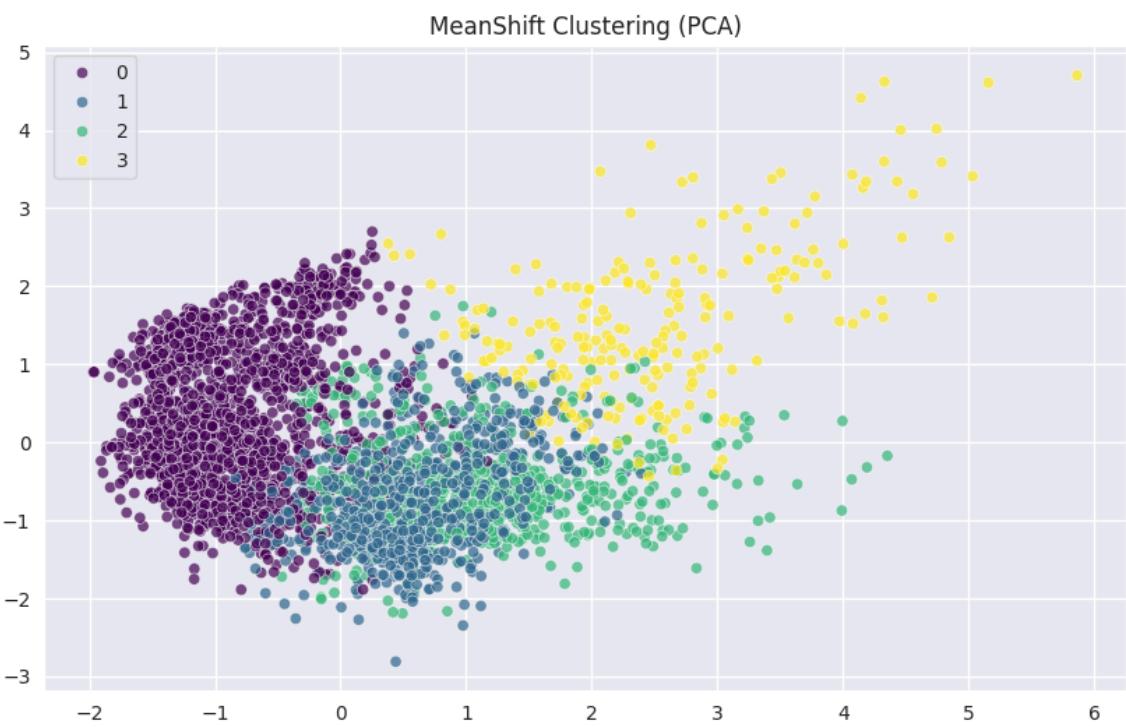
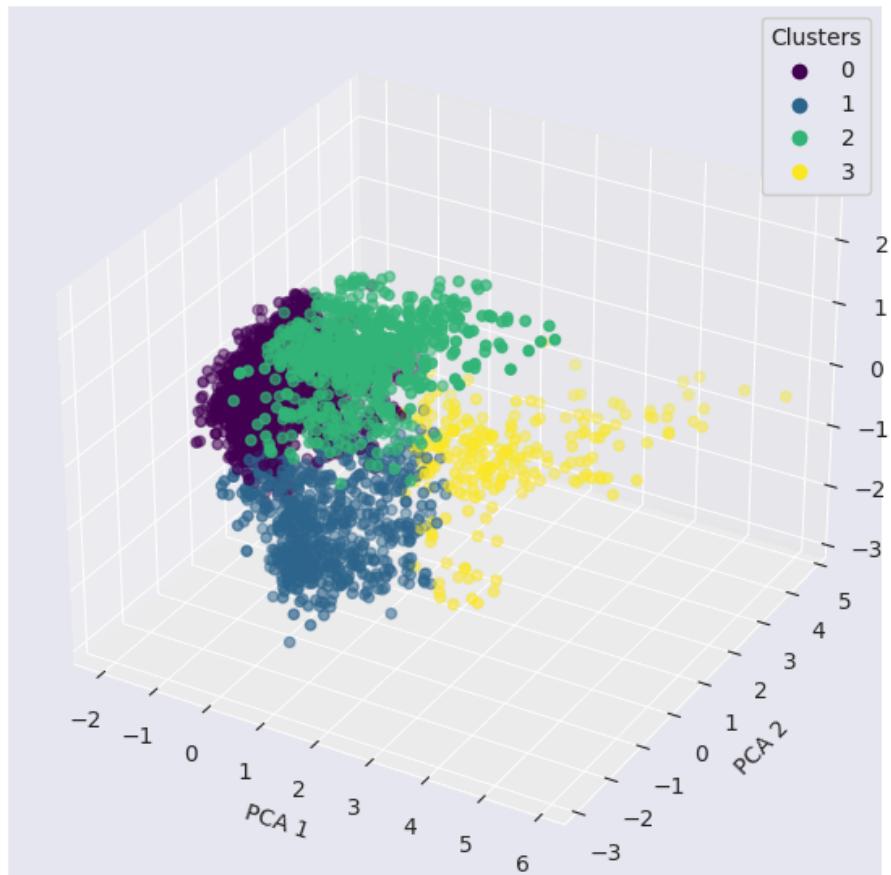


تحلیل:

- برای انتخاب تعداد مناسب خوش، معمولاً از روش برش دندوگرام استفاده می‌شود.
- **تعداد خوشها (Clusters):** با توجه به ارتفاع خطوط و نحوه ادغام خوشها، به نظر می‌رسد که با برش دندوگرام در سطح فاصله‌ای حدود 30 تا 50 داده‌ها به 6 خوشی مجزا تقسیم می‌شوند
- دندوگرام به خوبی نشان می‌دهد که برخی خوشها (مثلاً سمت چپ) نسبتاً سریع‌تر (در سطوح پایین‌تر) ادغام شده‌اند، به این معنی که اعضای آن خوشها شباهت بیشتری به هم دارند.
- در مقابل، برخی خوشها (سمت راست نمودار) در سطوح بالاتری ادغام شده‌اند، که نشان‌دهنده پراکندگی بیشتر در آن داده‌هاست

MeanShift 3.2.4

- **روش و برآورد پارامتر:**
الگوریتم MeanShift بدون نیاز به تعیین تعداد خوش از پیش عمل می‌کند. در این روش ابتدا مقدار مناسب برای `bandwidth` با استفاده از تابع `estimate_bandwidth` تعیین می‌شود.
- **مزایا:**
MeanShift به صورت خودکار خوشها را شناسایی کرده و برای داده‌هایی که ساختارهای پیچیده‌تری دارند کاربرد دارد.



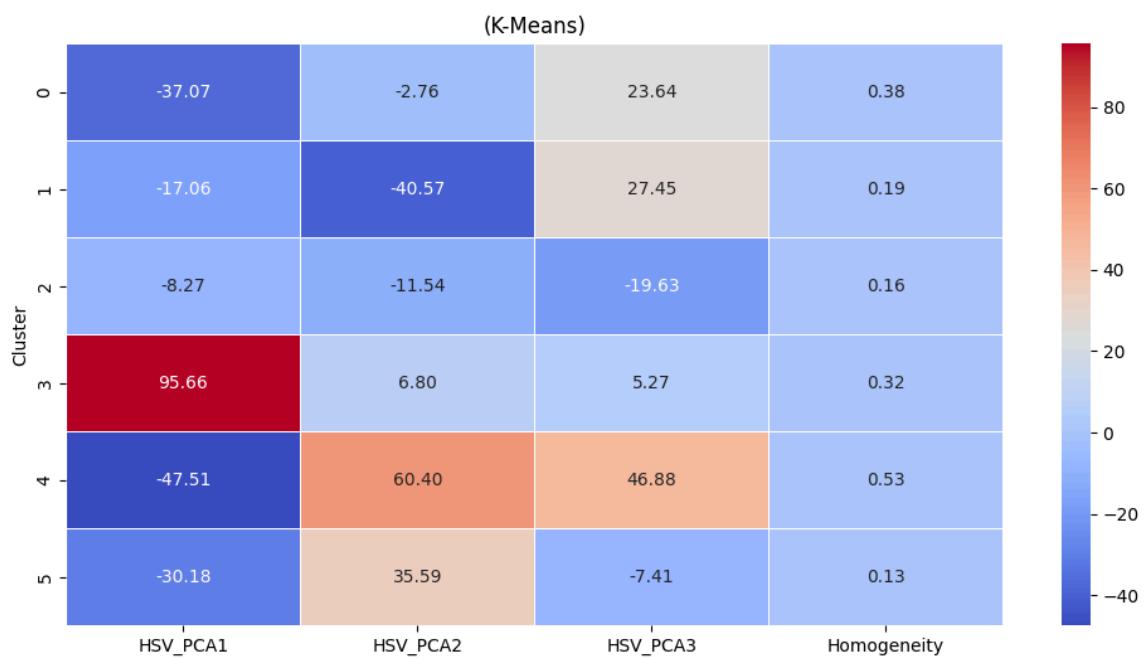
خروجی:

quantile=0.08

Estimated Bandwidth: 1.22

MeanShift - Silhouette Score: 0.3904

heatmap 3.3



بر اساس نمودار Heatmap، ویژگی‌های میانگین و همگنی خوش‌ها به شرح زیر خلاصه می‌شود:

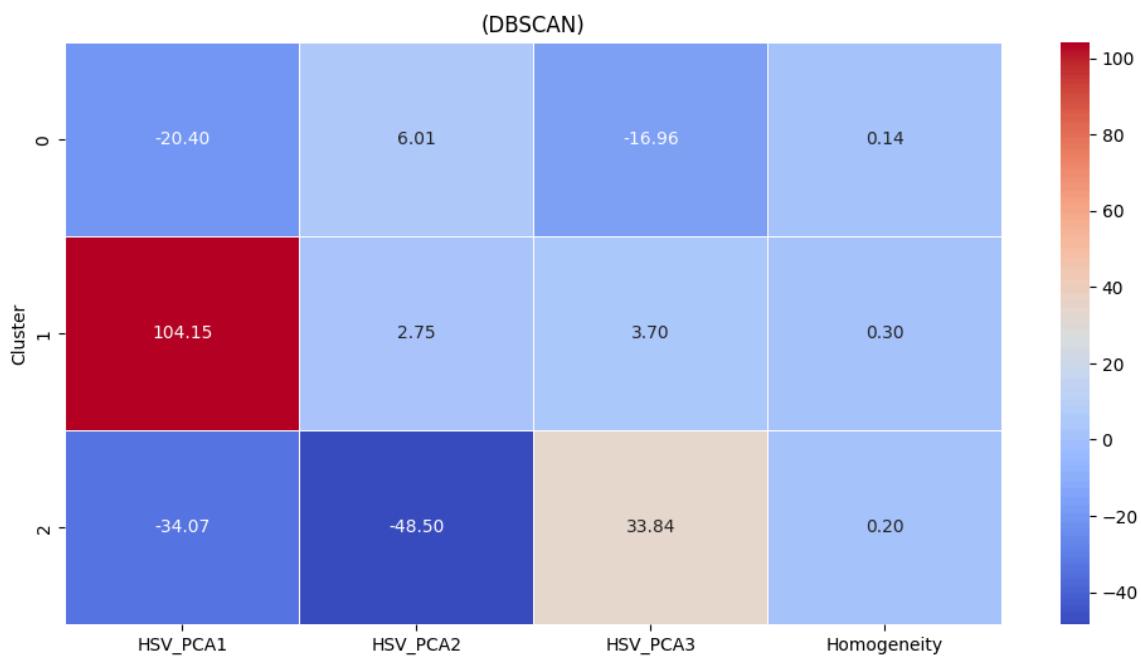
- خوش‌هه ۴ دارای بالاترین مقدار همگنی بوده و از نظر ویژگی‌های PCA2 و PCA3 نیز متمایز است. این خوش‌هه از نظر کیفیت خوشنامندی، بهترین عملکرد را دارد.

- خوشه ۳ با مقدار بسیار بالای (95.66) HSV_PCA1 از نظر ویژگی‌ها متمایز است، اما همگنی آن در سطح متوسط (0.32) قرار دارد.

- خوشه‌های ۱، ۲ و ۵ دارای مقادیر پایین‌تری از همگنی بوده (زیر ۰.۲) و تمایز مشخصی از نظر مؤلفه‌های PCA ندارند.

- خوشه ۰ همگنی قابل قبولی (0.38) دارد و دارای ترکیبی از ویژگی‌های مثبت و منفی در مؤلفه‌های

در مجموع، خوشه ۴ از نظر کیفیت تفکیک خوشه‌ای، مطلوب‌ترین عملکرد را نشان می‌دهد.

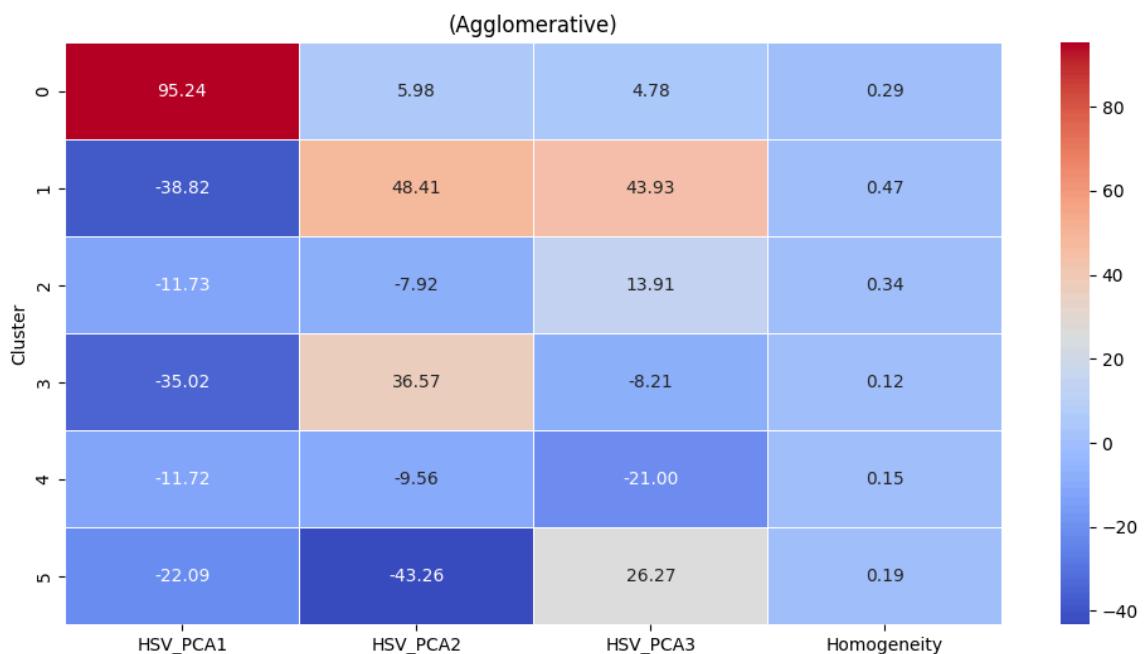


- خوشه ۱ بیشترین مقدار در HSV_PCA1 (104.15) را دارد و از نظر همگنی (0.30) نیز بهترین عملکرد را بین خوشه‌ها ارائه می‌دهد.

- خوشه ۲ دارای بیشترین مقدار در HSV_PCA2 (-48.50) و کمترین مقدار در HSV_PCA3 (33.84) است، اما همگنی پایینی دارد (0.20).

- خوشه ۰ از نظر مقادیر PCA ویژگی بارزی ندارد و همگنی آن نیز پایین است (0.14).

در مجموع، خوشه ۱ از نظر تمایز ویژگی و کیفیت خوشنده‌بندی نسبت به سایر خوشه‌ها عملکرد بهتری دارد، اما در مقایسه با K-Means، مقادیر همگنی در DBSCAN به طور کلی پایین‌تر هستند.



خوشه ۰ دارای مقدار بسیار بالای HSV_PCA1 و همگنی متوسط است. خوشه ۱ در HSV_PCA2 و HSV_PCA3 مقدار بالایی دارد و همگنی آن نیز نسبتاً بیشتر از بقیه خوشه‌هاست (۴۷٪). خوشه‌های دیگر ویژگی‌های متنوعی دارند، ولی همگنی کمتری نسبت به خوشه ۱ دارند.



خوشه ۲ دارای بالاترین مقدار HSV_PCA1 است، که نشان می‌دهد این خوشه احتمالاً گروهی متمایز با ویژگی رنگی قوی در بُعد اول است.

خوشه ۳ همگن‌ترین خوشه است ($\text{Homogeneity} = 0.48$) و همچنین در PCA2 و PCA3 مقادیر نسبتاً بالایی دارد.

خوشه ۰ و خوشه ۱ نسبتاً پراکندگی بیشتری دارند و همگنی کمتری.

Evaluation : ۵ فاز

(دقت) Precision ◆

نشان می‌دهد از بین نمونه‌هایی که در یک خوشه قرار گرفته‌اند، چند درصد واقعاً باید با هم می‌بودند.

بالا بودن دقت = خوشه‌ها "حالص‌تر" هستند.

(بازخوانی) Recall ◆

نشان می‌دهد از بین همه نمونه‌هایی که باید با هم باشند، چند درصد واقعاً در یک خوشه قرار گرفته‌اند.

- بالا بودن بازخوانی = الگوریتم توانسته اغلب اعضای یک کلاس واقعی رو کنار هم قرار بده.

(میانگین موزون دقت و بازخوانی) F1-Score ◆

یک معیار کلی برای ترکیب دقت و بازخوانی. عددی بین ۰ تا ۱ که تعادلی از هر دو رو نشون میده.

- مفید برای مقایسه نهایی الگوریتم‌ها.
- Agglomerative با 0.5951 عملکرد کلی بهتری نسبت به بقیه داشته.

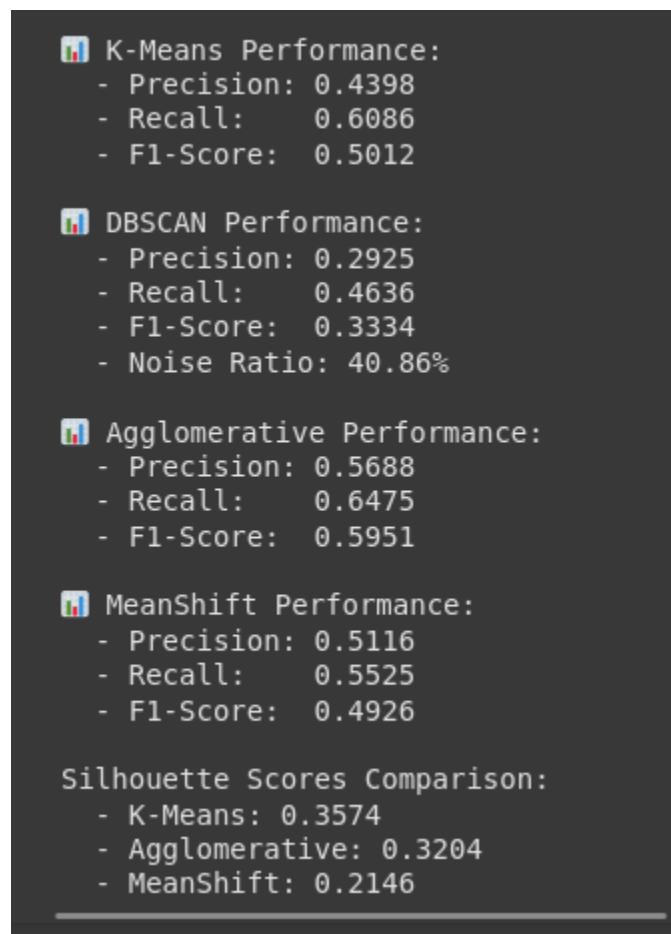
DBSCAN (نسبت نویز) – فقط برای Noise Ratio ◆

نشان می‌دهد چند درصد داده‌ها به هیچ خوش‌های تعلق نگرفته‌اند.

Silhouette Score ◆

شاخصی بین -1 تا +1 برای ارزیابی کیفیت خوش‌بندی:

- نزدیک ۱: خوش‌ها کاملاً مجزا.
- نزدیک ۰: خوش‌ها در هم تنیده.
- منفی: نمونه‌ها بیشتر به خوش‌های دیگر تعلق دارند.



K-Means ◆

تحلیل: با وجود دقت نسبتاً پایین، K-Means دارای بالاترین Silhouette Score است، که نشان‌دهنده مرزبندی واضح‌تری بین خوش‌ها است. این الگوریتم خوش‌هایی با تقسیک بهتر اما خلوص کمتر ایجاد کرد.

است.

DBSCAN ◆

تحلیل: عملکرد ضعیفتر در مقایسه با سایر الگوریتم‌ها. همچنین درصد بالای نقاط نویز (نژدیک به ۴۱٪) نشان‌دهنده حساسیت زیاد الگوریتم به پارامترها و ساختار داده‌هاست.

Agglomerative Clustering ◆

تحلیل: بهترین عملکرد کلی را از نظر دقت، بازنگرانی و F1-Score دارد. این الگوریتم موفق شده است خوش‌هایی با ترکیب خوب از خلوص و پوشش ایجاد کند. ساختار سلسله‌مراتبی آن به تشخیص بهتر روابط بین داده‌ها کمک کرده است.

MeanShift ◆

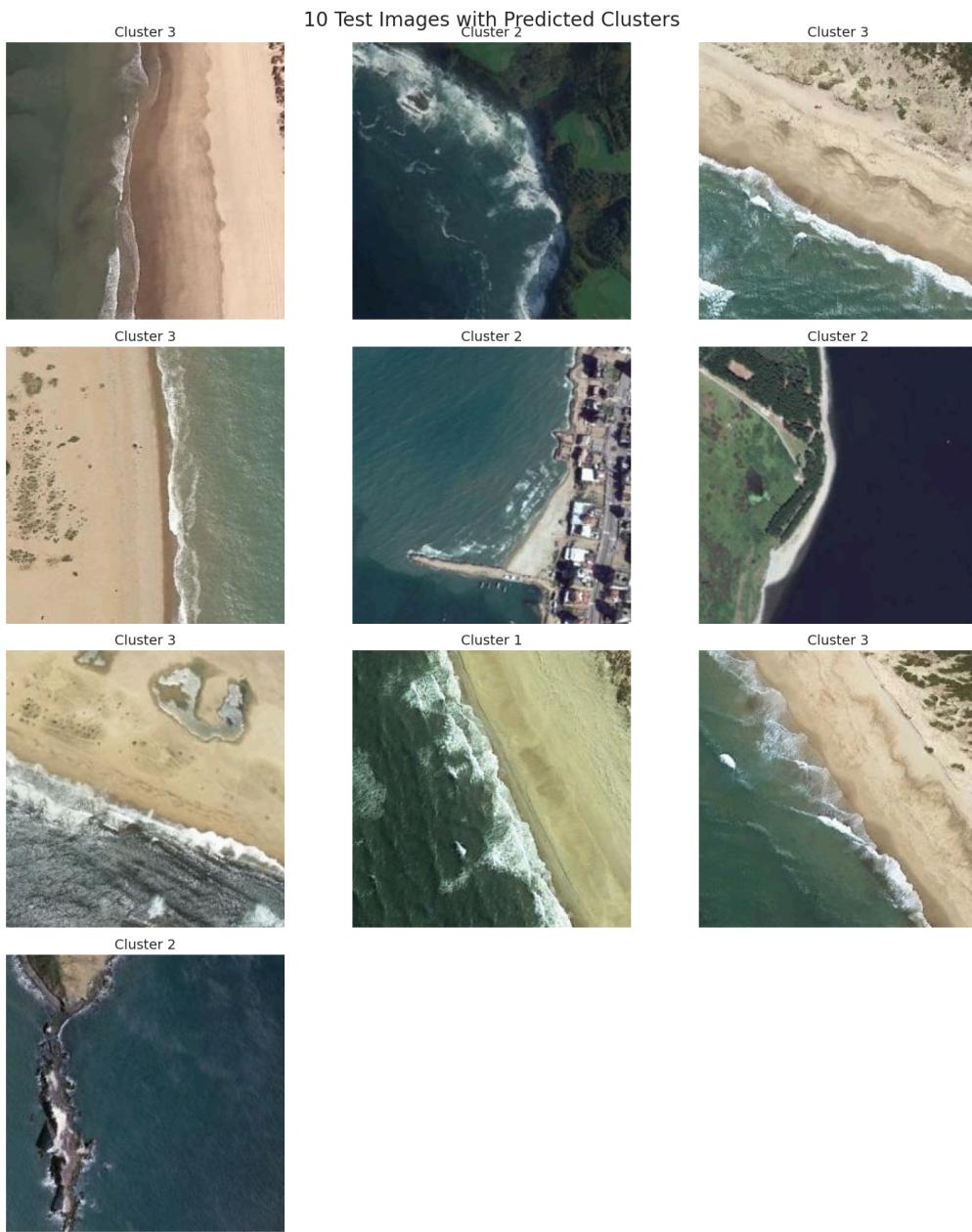
تحلیل: عملکرد متوسط در همه شاخص‌ها. اگرچه خوش‌های تا حدودی معنادار هستند، ولی Silhouette Score پایین‌تر از سایر الگوریتم‌ها نشان‌دهنده همپوشانی بیشتر بین خوش‌هایی است.

نتیجه نهایی:

الگوریتم Agglomerative Clustering با توجه به بیشترین مقدار F1-Score و توازن مناسب بین Precision و Recall، بهترین عملکرد کلی را در این مسئله خوشبینی داشته است. الگوریتم K-Means نیز با بالاترین Silhouette Score، از منظر کیفیت تقسیم خوش‌های عملکرد خوبی دارد. در مقابل، DBSCAN به دلیل تعداد بالای نقاط نویز، عملکرد قابل قبولی از خود نشان نداده است.

فاز 6 : Prediction

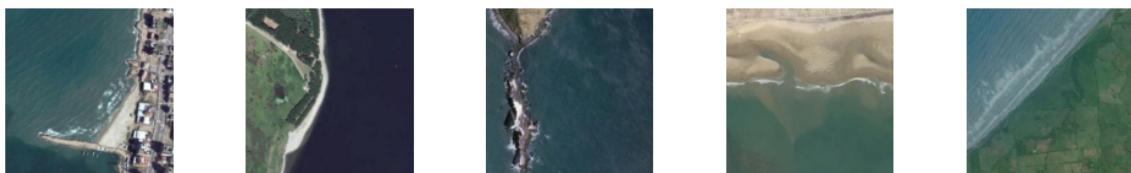
در فاز 6، تصاویر تست بارگذاری و ویژگی‌های آن‌ها استخراج می‌شود. سپس با استفاده از PCA بعد ویژگی‌های رنگی کاهش می‌یابد و ویژگی‌های دیگر نیز در کنار آن‌ها قرار می‌گیرند. پس از انتخاب ویژگی‌های مناسب، داده‌ها نرم‌السازی می‌شوند و با استفاده از مدل KMeans پیش‌بینی خوش‌ها انجام می‌شود. نتایج پیش‌بینی در یک فایل CSV ذخیره و 10 تصویر اول همراه با خوش‌های پیش‌بینی‌شده نمایش داده می‌شوند. همچنین، 5 تصویر مشابه از همان خوش برای هر تصویر تست نمایش داده می‌شود.



Cluster 3 - Related Samples to Test Image 1



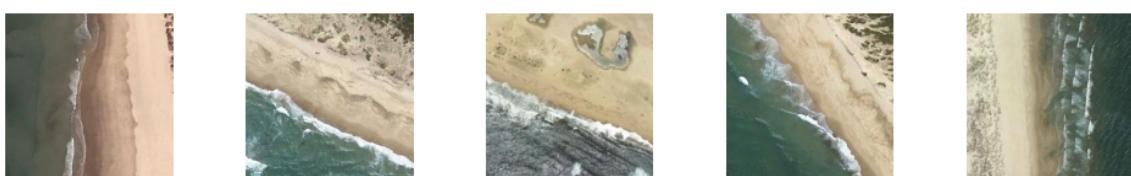
Cluster 2 - Related Samples to Test Image 2



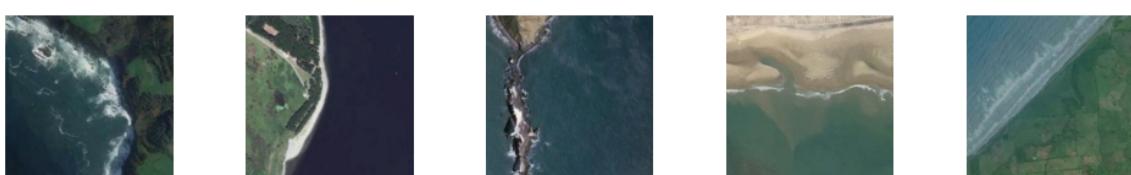
Cluster 3 - Related Samples to Test Image 3



Cluster 3 - Related Samples to Test Image 4



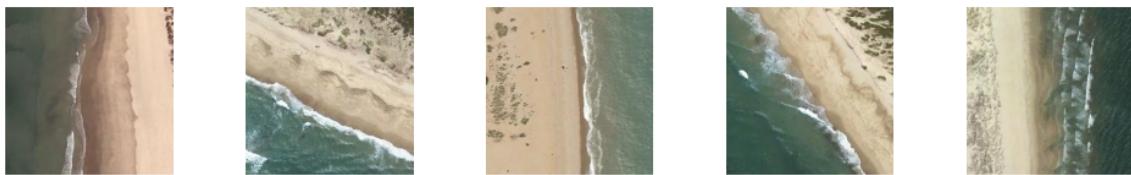
Cluster 2 - Related Samples to Test Image 5



Cluster 2 - Related Samples to Test Image 6



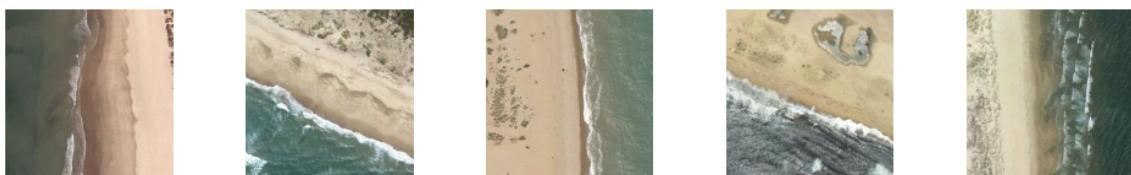
Cluster 3 - Related Samples to Test Image 7



Cluster 1 - Related Samples to Test Image 8



Cluster 3 - Related Samples to Test Image 9



Cluster 2 - Related Samples to Test Image 10

