



دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر



گزارش پروژه‌ی نهایی

درس یادگیری ماشین

شماره دانشجویی

810199584

810100377

810101229

810101183

نام و نام خانوادگی

مهرسا همت پناه

تابان سلیمانی

سارا عطار صادق صومعه سرایی

سارا سام پور

تابستان 1402

فهرست

4.....	چکیده.....
5.....	معرفی داده ها.....
6.....	بخش 1 - تمیز کردن داده و استخراج ویژگی.....
6.....	ویژگی های تشخیص نویز.....
7.....	عدم شفافیت.....
9.....	مکانیزم تشخیص لبه.....
11.....	میزان روشنایی و توزیع رنگ ها.....
16.....	بافت.....
19.....	بخش 2 - طبقه بندی.....
19.....	2.1-اموزش مدل برای ویژگی های داده شده.....
19.....	درخت تصمی:.....
20.....	رگرسیون لجستیک.....
21.....	SVM.....
23.....	Random Forest.....
25.....	2.2-اموزش مدل برای ویژگی های استخراج شده.....
25.....	درخت تصمیم.....
26.....	رگرسیون لجستیک.....
27.....	SVM.....
28.....	Random Forest.....
29.....	شبکه عصبی.....
32.....	بخش 3 - خوشابندی.....
32.....	3.1-مدل خوشابند برای ویژگی های داده شده.....
32.....	GMM.....
42.....	Hierarchical.....

52.....	k-means
59.....	3.2 مدل برای ویژگی های استخراج شده
59.....	GMM
66.....	Hierarchical
73.....	k-means

چکیده

در این پژوهه قصد داریم مدل هایی برای طبقه بندی و خوشه بندی عکس ها بر اساس واقعی یا مصنوعی بودن آن ها طراحی کنیم. هدف از این کار ارتقای کیفیت عکس های مصنوعی تولید شده توسط هوش مصنوعی می باشد. البته نگرانی های امنیتی و حفظ حریم خصوصی نیز مطرح است. یافتن تکنیک های موثر برای تشخیص تصاویر جعلی بسیار ضروریست چرا که اگر مطالب جعلی این چنینی به جای واقعیت منتشر شود، به اعتبار انسان ها و موسسات لطمه می زند و نظم اجتماعی را به هم می ریزد.

در ابتدا برای این کار به جمع آوری داده پرداختیم. داده های جمع آوری شده متشکل از عکس های واقعی و مصنوعی در سه گروه دریا، کوه و جنگل هستند. در ادامه به معرفی ویژگی های داده ها برای استفاده در مدل و الگوریتم یادگیری ماشین نهایی می پردازیم.

معرفی داده ها

برای این پروژه، هر فرد 30 عکس واقعی در سه کلاس 10 تایی کوه، دریا و جنگل جمع آوری کرده است. سعی بر آن بوده است که عکسها از تنوع خوبی برخوردار باشند. در مرحله دوم نیز هر فرد 30 عکس مصنوعی توسط ابزارهای ذکر شده گردآوری کرده است. در مجموعه عکس های مصنوعی، برای عمل یادگیری بهتر از عکسهایی با فرمت انیمیشنی نیز استفاده شده است.

● عکس های مصنوعی

همانطوری که در صورت پروژه ذکر شده، برای تولید عکس های مصنوعی در گروه های خواسته شده، از الگوریتم های مبتنی بر شبکه های مولد متخصص¹ استفاده کردیم.

● عکس های واقعی

عکس های واقعی در سه گروه دریا، کوه و جنگل هستند.

¹ Generative adversarial network(GAN)

بخش 1 - تمیز کردن داده و استخراج ویژگی

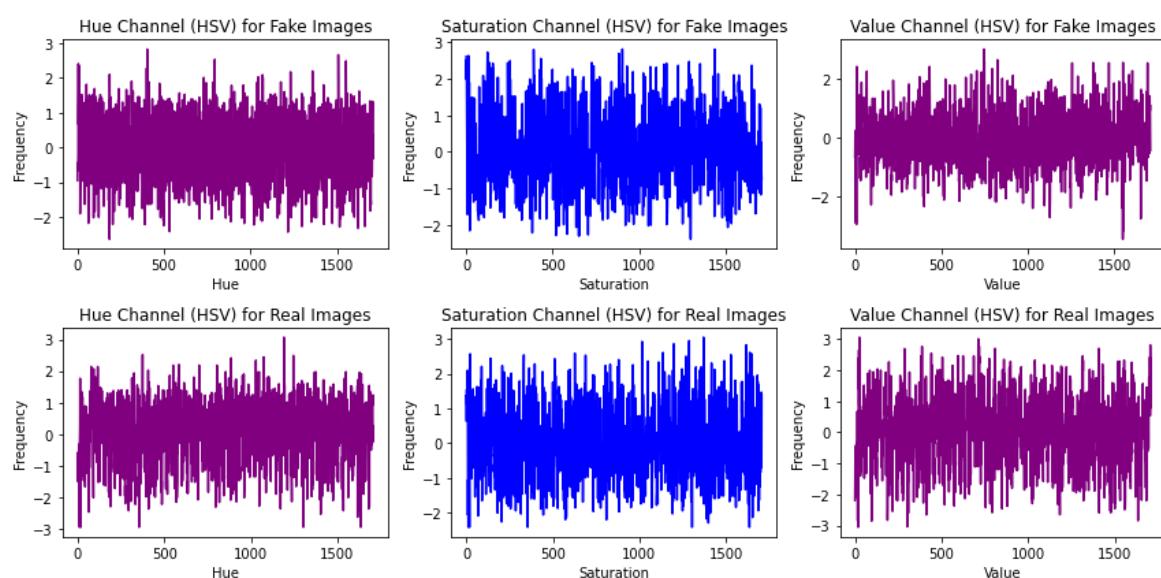
استخراج ویژگی بخشی از پروسه‌ی کاهش ابعاد می‌باشد که در آن دسته‌ای از دیتای خام به گروههای قابل مدیریت و کوچکتر تقسیم می‌شود. به زبان ساده، هریک از پیکسل‌های یک عکس بخشی از دیتا می‌باشد و پردازش تصاویر استخراج ویژگی‌های مفید یک عکس با کاهش ابعاد و حفظ پیکسل‌هایی است که ویژگی‌های عکس را توصیف می‌نمایند.

ویژگی‌های تشخیص نویز

برای تشخیص نویز از ویژگی‌هایی مانند طول موج^۱، میزان اشباع رنگ‌ها^۲ و شدت رنگ^۳ در عکس می‌توان استفاده کرد.

- Hue : طول موج غالب در عکس را مشخص می‌کند و یک کanal تعیین کننده برای رنگ است.
- Saturation : بیانگر میزان اشباع (purity / shades) طول موج / رنگ در عکس می‌باشد.
- Value : شدت رنگ را توصیف می‌کند.

شکل 1 میانگین سه ویژگی Hue/Saturation/Value برای دو کلاس fake و real در عکس‌ها را نشان می‌دهد.

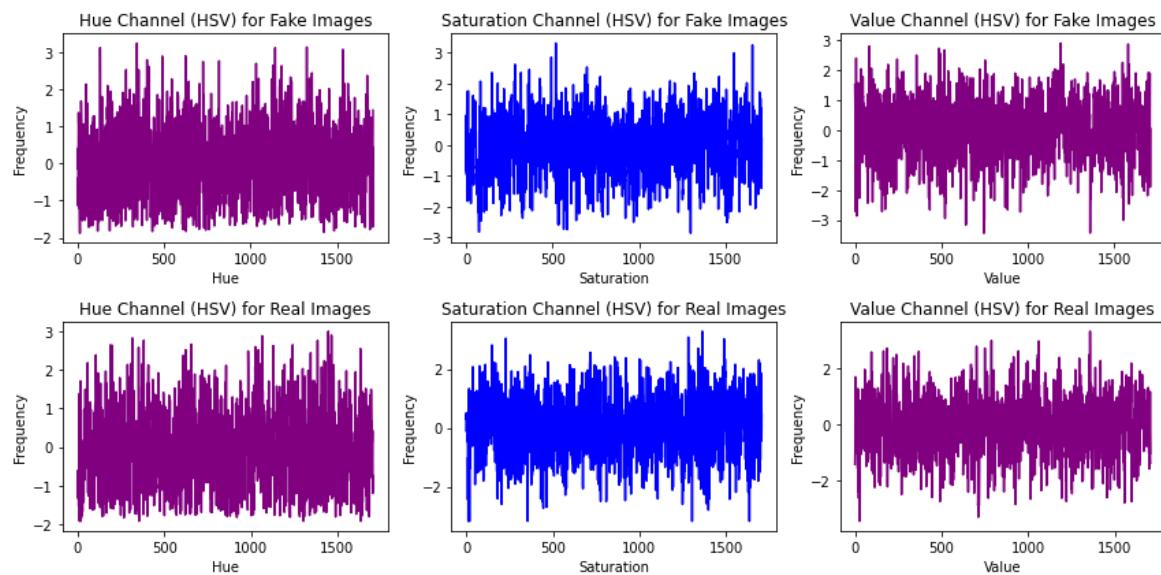


شکل 1 میانگین مقادیر ویژگی‌های Hue/Saturation/Value برای عکس‌های واقعی و غیر واقعی

Wave Length¹
Saturation²
Intensity³

همانطوری که در شکل 1 مشاهده می‌کنید، شدت رنگ در عکس‌های واقعی نسبت به غیرواقعی به طور میانگین، بیشتر است. یا به عنوان مثال، میزان اشباع رنگ و نیز طول موج غالب در عکس‌های غیرواقعی نسبت به واقعی بیشتر است.

شکل 2 انحراف معیار سه ویژگی Hue/Saturation/Value برای دو کلاس real و fake، در عکس‌ها را نشان می‌دهد که عملاً میزان انحراف معیار مقادیر گزارش شده برای ویژگی‌های اشباع و شدت رنگ در عکس‌های واقعی نسبت به غیرواقعی کمتر است.

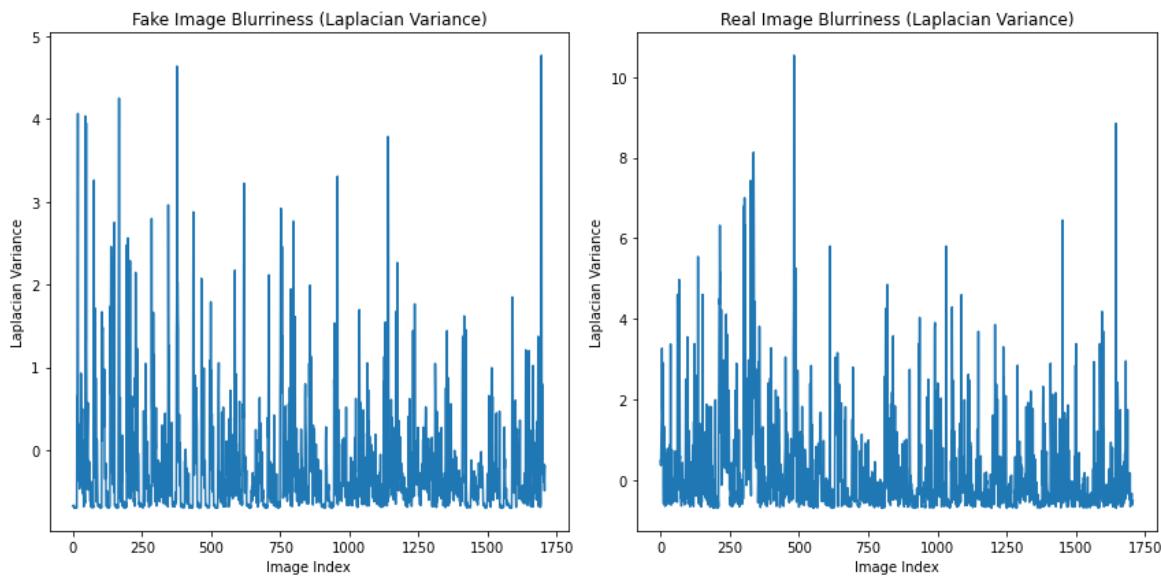


شکل 2 انحراف معیار مقادیر ویژگی‌های Hue/Saturation/Value برای عکس‌های واقعی و غیر واقعی

عدم شفافیت^۱

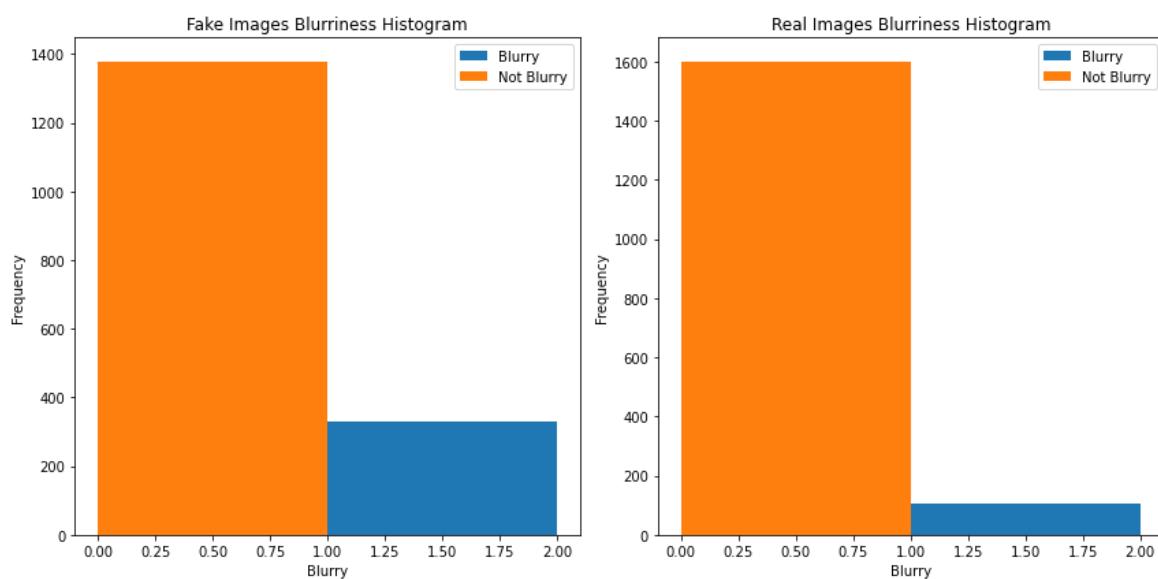
ابتدا با استفاده از تابع Laplacian در cv2، فاصله‌ی کانونی در هر عکس محاسبه شده و سپس از واریانس مقادیر به دست آمده، به عنوان متريک تعیین کننده‌ی عدم شفافیت استفاده شده است. شکل 3 مقادیر واریانس Laplacian را برای دو کلاس عکس‌های واقعی و غیرواقعی نشان می‌دهد. همانطوری که در شکل 3 مشاهده می‌کنید، میزان واریانس Laplacian در عکس‌های واقعی نسبت به عکس‌های غیرواقعی بیشتر است.

¹ Blurriness



شکل 3 میزان واریانس Laplacian برای عکس‌های واقعی و غیرواقعی

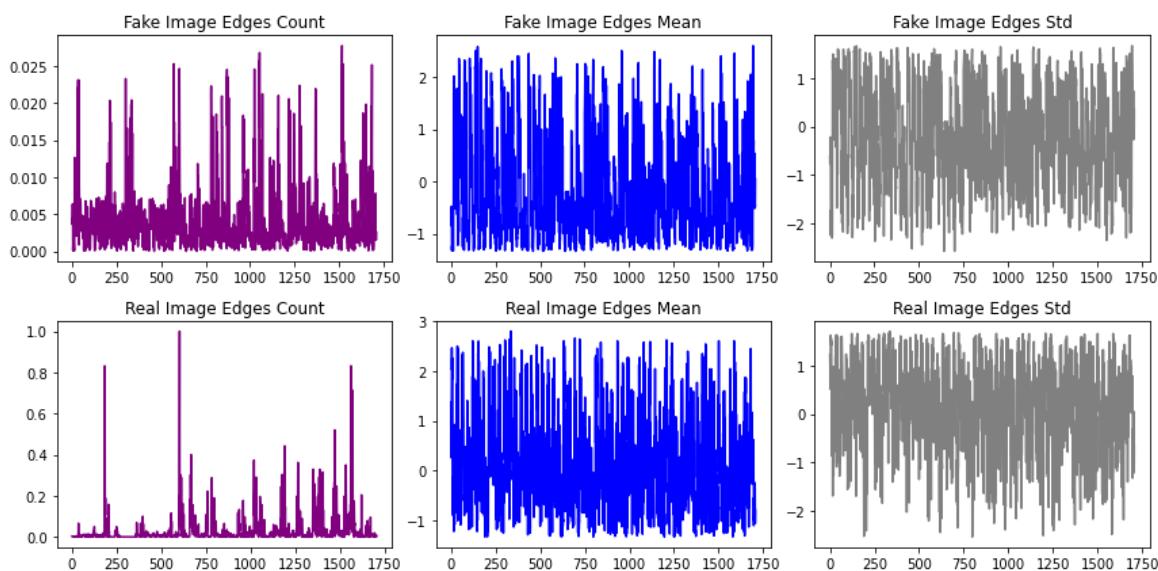
شکل 4 هیستوگرام توزیع عدم شفافیت در عکس‌های واقعی و غیرواقعی را (با تعیین threshold = 100 به دست آمده در قسمت قبل) نشان می‌دهد.



شکل 4 هیستوگرام میزان عدم شفافیت در عکس‌های واقعی و غیرواقعی

مکانیزم تشخیص لبه^۱

ابتدا با استفاده از تابع cvtColor در cv2، هر عکس به مقیاس خاکستری آن تبدیل می‌شود. سپس با فراخوانی متدهای آن مشخص می‌شود. مقادیر minVal و maxVal نیز در این متدهای ترتیب 100 و 200 در نظر گرفته شده است. در ادامه نیز تعداد، میانگین و انحراف معیار لبه‌ها برای عکس‌های واقعی و غیرواقعی در شکل ۵ گزارش شده است و سه پارامتر به دست آمده به عنوان ویژگی‌های استخراج شده در این بخش درنظر گرفته شده‌اند.

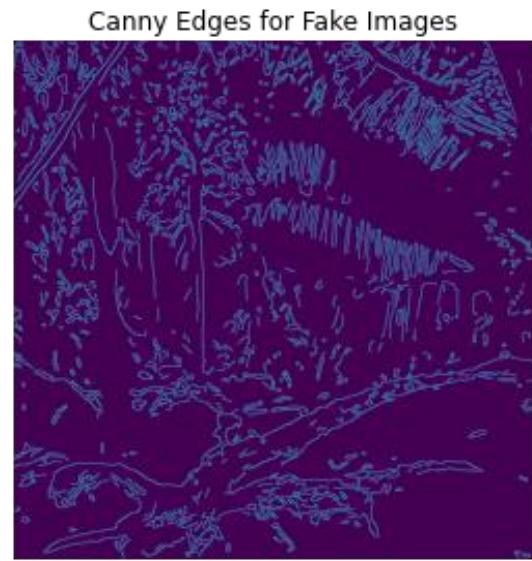


شکل ۵ تعداد، میانگین و انحراف معیار لبه در عکس‌های واقعی و غیرواقعی

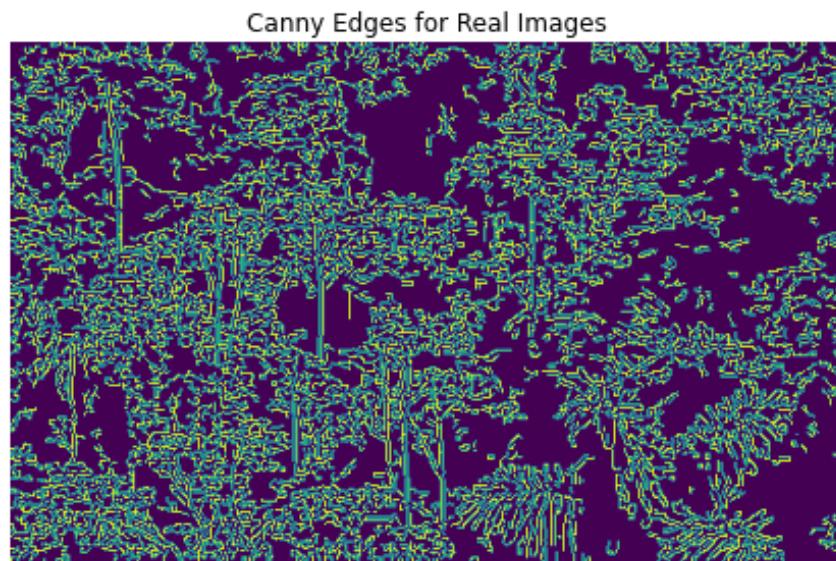
همانطوری که در شکل ۵ مشاهده می‌کنید، به طور میانگین تعداد لبه‌های تشخیص داده شده در بین عکس‌های واقعی بیشتر بوده است. (این موضوع به خوبی در عکس‌های ۶ و ۷ نیز مشهود است). به علاوه‌ی اینکه میانگین و انحراف معیار مقادیر پیکسل‌ها در لبه‌های به دست آمده در عکس‌های واقعی نسبت به غیرواقعی بیشتر است.

Canny Edge Detection^۱

شکل 6 لبه‌های تشخیص داده شده برای عکس غیرواقعی 150400011_fake_dall.e_jungle_1.jpg را نشان می‌دهد.



شکل 6 لبه‌های تشخیص داده شده برای اولین عکس غیرواقعی از کلاس جنگل در شکل 7 نیز لبه‌های تشخیص داده شده برای عکس واقعی 150400011_real_none_jungle_1.jpg مشاهده می‌کنید.



شکل 7 لبه‌های تشخیص داده شده برای اولین عکس واقعی از کلاس جنگل

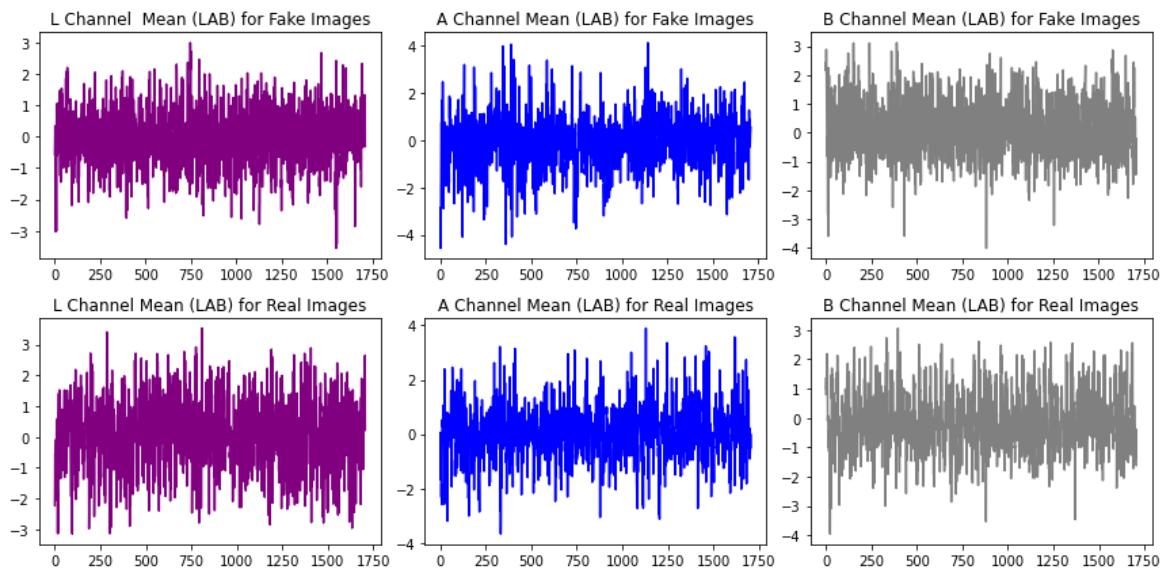
میزان روشنایی و توزیع رنگ‌ها

در این قسمت از فضای COLOR_BGR2LAB در cv2 برای توصیف عکس‌ها استفاده شده است.

- Lightning : پارامتر اول خروجی یا L، میزان روشنایی رنگ در عکس را نشان می‌دهد که با مقدار intensity قابل جایگزینی است.
- A : مولفه‌ی رنگ متغیر از سبز تا جوهر قرمز^۱
- B : مولفه‌ی رنگ متغیر از آبی تا زرد

شکل 8 میانگین سه ویژگی L/A/B برای دو کلاس fake و real، در عکس‌ها را نشان می‌دهد.

همانطوری که در شکل 8 مشاهده می‌کنید، روشنایی رنگ در عکس‌های واقعی نسبت به غیرواقعی به طور میانگین، کمتر بوده و نیز مقدار مولفه‌ی A و B به طور میانگین در عکس‌های غیرواقعی نسبت به واقعی بیشتر است.

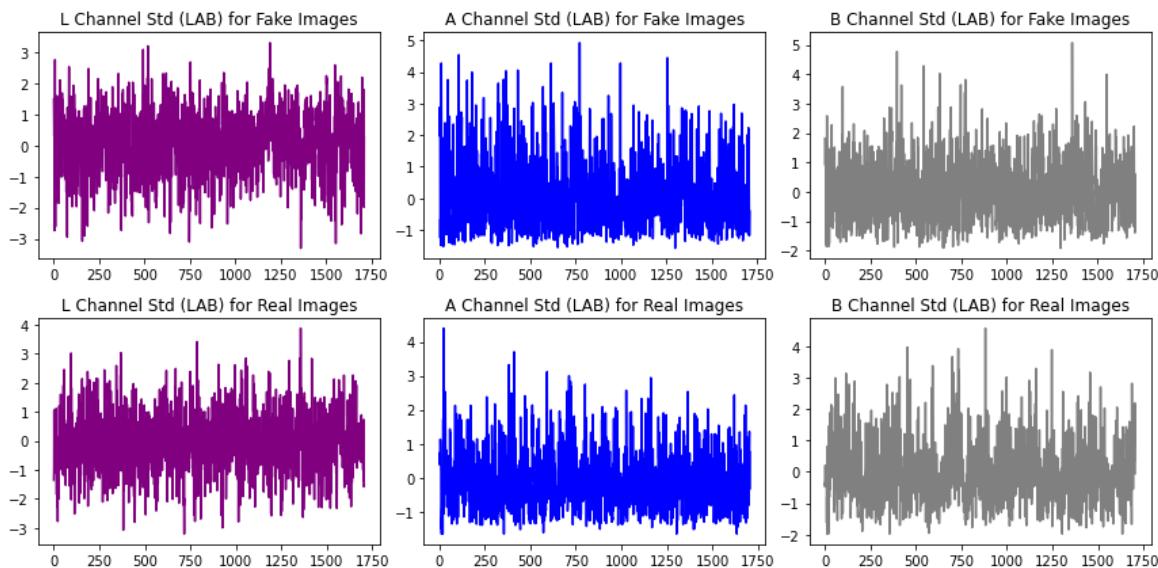


شکل 8 میانگین مقادیر ویژگی‌های L/A/B در عکس‌های واقعی و غیرواقعی

شکل 9 انحراف معیار سه ویژگی L/A/B برای دو کلاس fake و real، در عکس‌ها را نشان می‌دهد.

همانطوری که در شکل 9 مشاهده می‌کنید، انحراف معیار روشنایی رنگ در عکس‌های واقعی نسبت به غیرواقعی، بیشتر بوده و نیز انحراف معیار مقادیر مربوط به مولفه‌ی A و B در عکس‌های غیرواقعی نسبت به واقعی بیشتر است.

Magenta¹

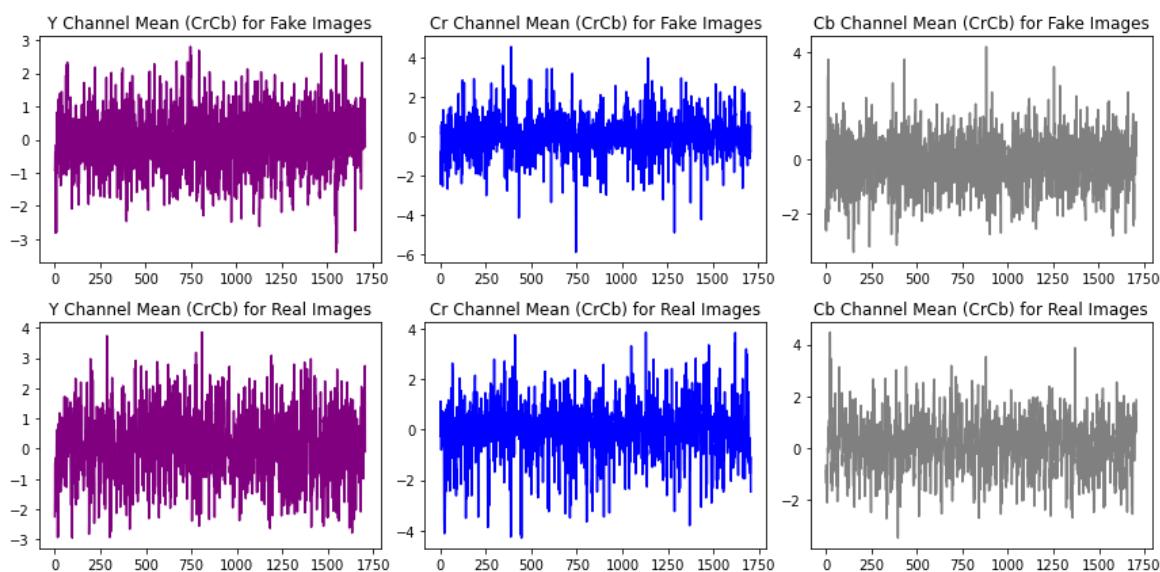


شکل 9 انحراف معيار مقادير ويزگي های L/A/B در عکس های واقعی و غیر واقعی

علاوه بر اين يك راه ديگر برای به دست آوردن توزيع رنگ، توصيف عکس در فضای COLOR_BGR2YCrCb cv2 می باشد.

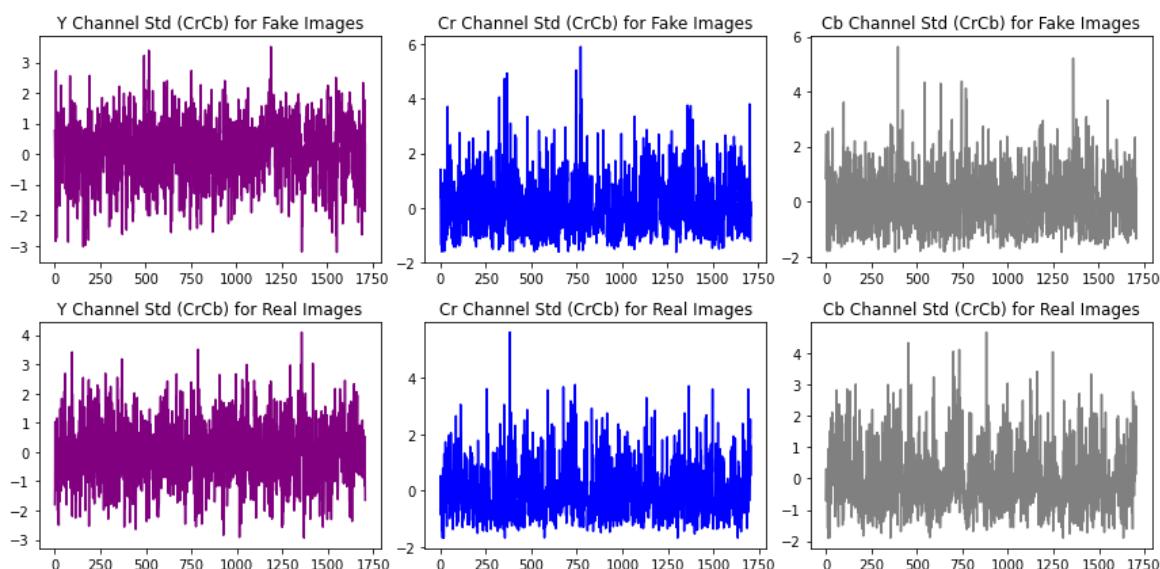
- Y : تشعشع به دست آمده از فضای رنگ RGB پس از تصحیح گاما
- Cr : بیانگر میزان تفاوت مولفه‌ی قرمز از تشعشع می باشد.
- Cb : بیانگر میزان تفاوت مولفه‌ی آبی از تشعشع می باشد.

شکل 10 ميانگين سه ويزگي Y/Cr/Cb برای دو کلاس fake و real در عکس‌ها را نشان می‌دهد. همانطوری که در شکل 10 مشاهده می‌کنید، میزان تشعشع یا مقدار پارامتر Y در عکس‌های واقعی نسبت به غیر واقعی به طور ميانگين، بيشتر بوده و نيز مقدار مولفه‌ی Cr و Cb به طور ميانگين در عکس‌های غیر واقعی نسبت به واقعی بيشتر است.



شکل 10 میانگین مقادیر ویژگی‌های $Y/Cr/Cb$ در عکس‌های واقعی و غیرواقعی

شکل 11 انحراف معیار سه ویژگی $Y/Cr/Cb$ برای دو کلاس fake و real در عکس‌ها را نشان می‌دهد. همانطوری که در شکل 11 مشاهده می‌کنید، میزان انحراف معیار تشعشع یا مقدار پارامتر Y در عکس‌های واقعی نسبت به غیرواقعی بیشتر بوده و نیز انحراف معیار مقدار مولفه‌ی Cr و Cb در عکس‌های غیرواقعی نسبت به واقعی بیشتر است.



شکل 11 انحراف معیار مقادیر ویژگی‌های $Y/Cr/Cb$ در عکس‌های واقعی و غیرواقعی

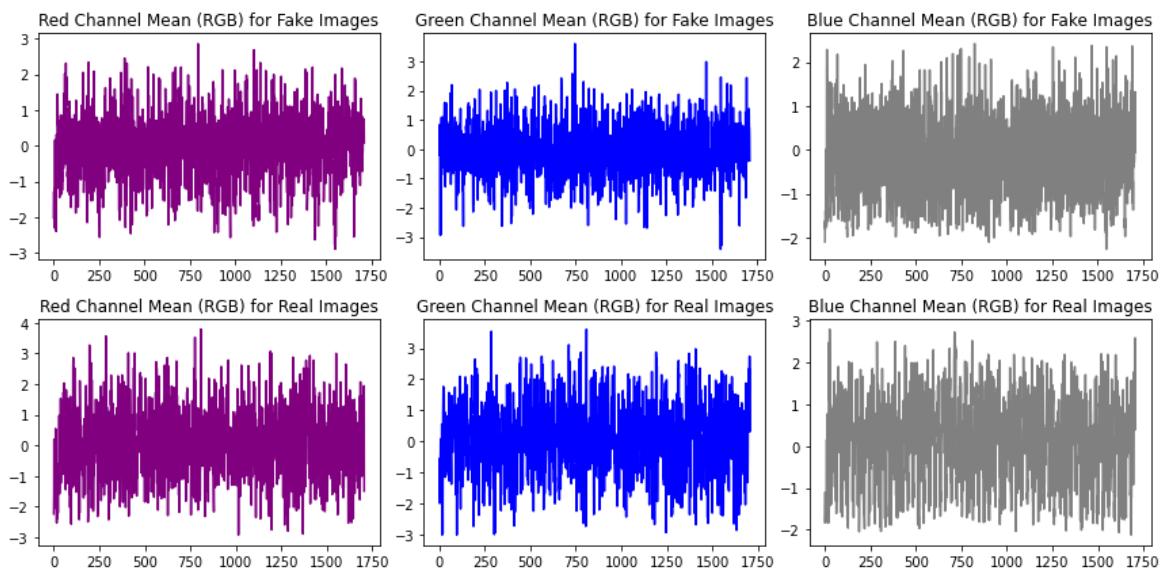
همانطوری که می‌دانید، سرراست‌ترین راه برای به دست آوردن توزیع رنگ، توصیف عکس در فضای COLOR_BGR2RGB در cv2 می‌باشد.

- R : بیانگر میزان نور قرمز

- G : بیانگر میزان نور سبز

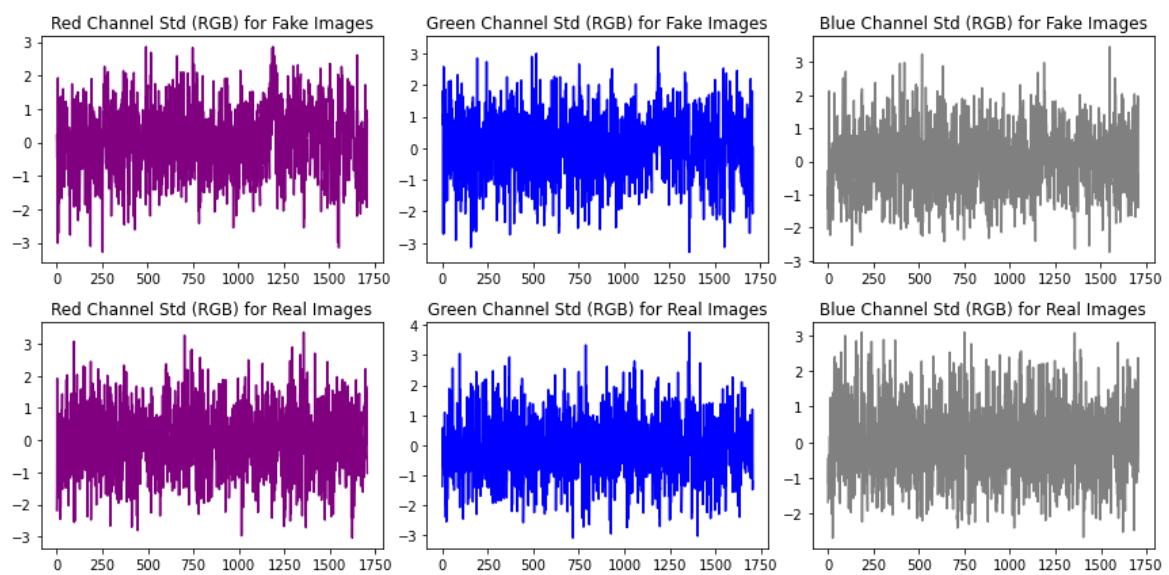
- B : بیانگر میزان نور آبی

شکل 12 میانگین سه ویژگی R/G/B برای دوکلاس fake و real، در عکس‌ها را نشان می‌دهد. همانطوری که در شکل 12 مشاهده می‌کنید، میزان نور قرمز، سبز و آبی در عکس‌های واقعی نسبت به غیرواقعی به طور میانگین، بیشتر است. (همانطوری که می‌دانید مقدار پیک نور آبی در عکس‌های مربوط کلاس دریا و مقدار پیک نور سبز در عکس‌های مربوط به کلاس جنگل رخ می‌دهد).



شکل 12 میانگین مقادیر ویژگی‌های R/G/B در عکس‌های واقعی و غیرواقعی

شکل 13 انحراف معیار مقادیر سه ویژگی R/G/B برای دوکلاس fake و real، در عکس‌ها را نشان می‌دهد. همانطوری که در شکل 13 مشاهده می‌کنید، انحراف معیار میزان نور قرمز، سبز در عکس‌های غیرواقعی نسبت به واقعی، بیشتر است و بالعکس میزان انحراف معیار نور آبی در عکس‌های واقعی نسبت به غیرواقعی بیشتر است.



شکل 13 انحراف معيار مقادير ويزگی های R/G/B در عکس های واقعی و غیر واقعی

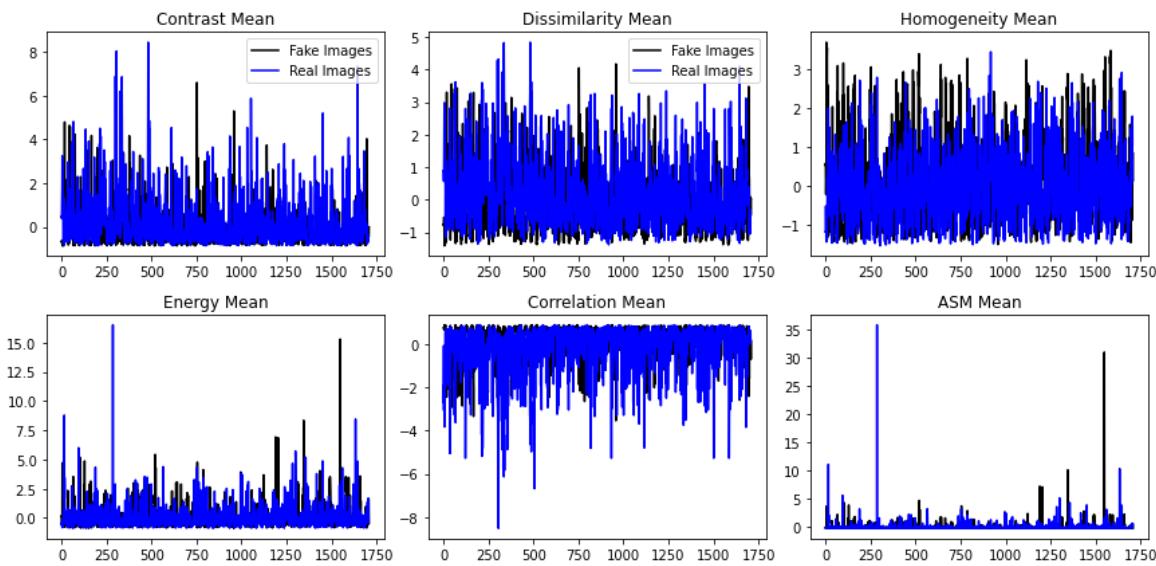
بافت

به صورت آماری، (Gray-level co-occurrence matrix(GLCM) روشی برای به دست آوردن بافت عکس می‌باشد که ارتباط مکانی بین پیکسل‌ها را در نظر می‌گیرد. این روش با محاسبه تعداد دفعات وقوع جفت پیکسل در یک تصویر با مقادیر خاص و در یک ارتباط فضایی مشخص (ایجاد ماتریس GLCM) کار می‌کند و سپس به استخراج معیارهای آماری از ماتریس به دست آمده می‌پردازد.

یک راه آسان برای به دست آوردن ماتریس GLCM استفاده از پکیج *scikit-image* می‌باشد که در ادامه توضیح مختصری در ارتباط با ویژگی‌های استخراج شده ارائه می‌گردد.

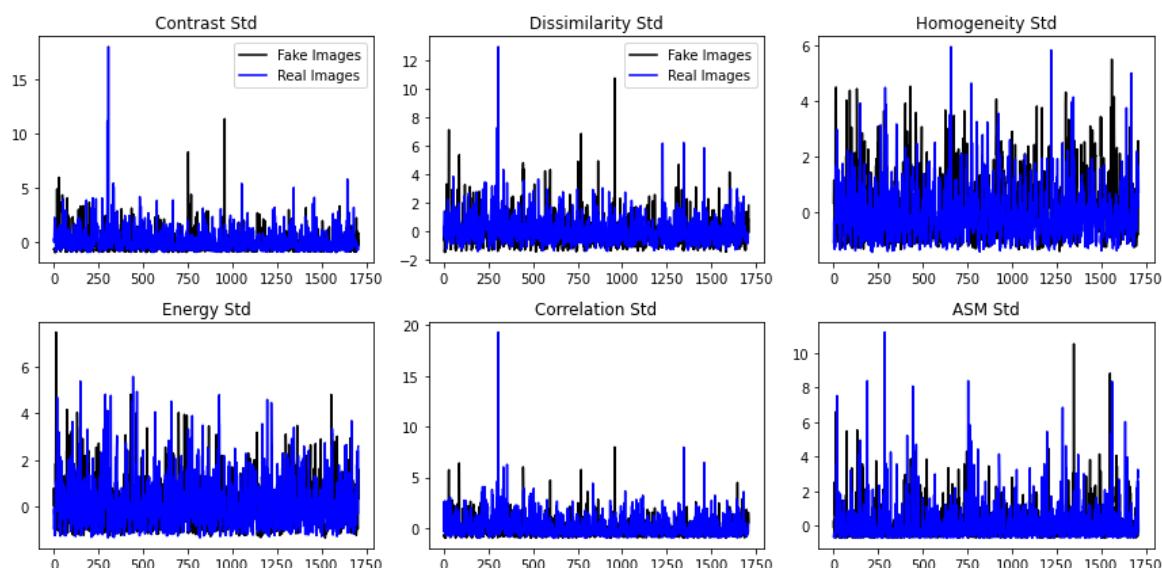
- Contrast : تغییرات محلی را در ماتریس Gray-level co-occurrence اندازه‌گیری می‌کند.
- Correlation : احتمال وقوع مشترک جفت پیکسل‌های مشخص شده را اندازه‌گیری می‌کند.
- Energy : جمع مجذور المان‌های ماتریس GLCM را اندازه‌گیری می‌کند که به عنوان پaramتر یکنواختی نیز شناخته می‌شود.
- Homogeneity : نزدیکی توزیع عناصر ماتریس GLCM تا ماتریس قطری GLCM را اندازه‌گیری می‌کند.

شکل 14 میانگین ویژگی‌های ASM، Energy، Correlation، Homogeneity، Dissimilarity و Contrast برای دو کلاس real و fake در عکس‌ها را نشان می‌دهد. همانطوری که در شکل 14 مشاهده می‌کنید، میزان Dissimilarity در عکس‌های واقعی نسبت به غیرواقعی به طور میانگین، بیشتر است و نیز میزان Contrast، Energy، Correlation، ASM، Homogeneity در عکس‌های واقعی نسبت به غیرواقعی به طور میانگین، کمتر است.



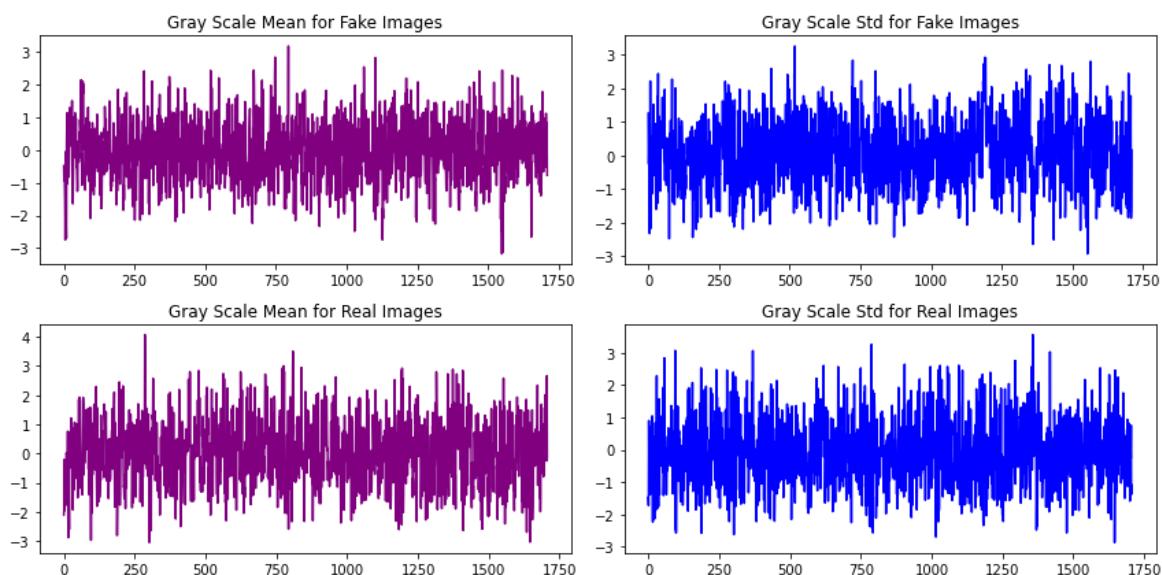
شکل 14 میانگین مقادیر ویژگی‌های Contrast/Dissimilarity/Homogeneity/Energy/Correlation/ASM در عکس‌های واقعی و غیرواقعی

شکل 15 انحراف معیار مقادیر ویژگی‌های ASM, Homogeneity, Energy, Correlation, Contrast در عکس‌ها را نشان می‌دهد. همانطوری که در شکل 15 Dissimilarity مشاهده می‌کنید، میزان انحراف معیار ASM, Dissimilarity, Contrast در عکس‌های واقعی نسبت به غیرواقعی بیشتر است و نیز میزان انحراف معیار مقادیر Homogeneity و Energy در عکس‌های واقعی نسبت به غیرواقعی کمتر است.



شکل 15 انحراف معیار مقادیر ویژگی‌های Contrast/Dissimilarity/Homogeneity/Energy/Correlation/ASM در عکس‌های واقعی و غیرواقعی

در شکل 16 نیز مقادیر مربوط به میانگین و انحراف معیار پیکسل‌های هر عکس در فضای Gray-scale را مشاهده می‌کنید. همانطوری که در شکل 16 مشاهده می‌کنید، مقدار میانگین و انحراف معیار پیکسل‌ها در مقیاس خاکستری در بین عکس‌های واقعی نسبت به غیرواقعی بیشتر است.



شکل 16 میانگین و انحراف معیار مقادیر ویژگی Gray-scale در عکس‌های واقعی و غیرواقعی

بخش 2 - طبقه‌بندی

2.1- آموزش مدل برای ویژگی‌های داده شده

درخت تصمیم:

درخت تصمیم یک الگوریتم است که بر اساس میزان انتروپی کار می‌کند به این صورت که ویژگی‌ها به صورت رئوس درخت انتخاب می‌شوند و براساس آن‌ها به کلاس‌های مختلف تقسیم می‌شوند عمق هر ویژگی بر اساس خلوصی که در برگ‌هایش ایجاد می‌کند بستگی دارد. این الگوریتم ساده و قابل فهم است ولی معمولاً دچار بیش برآذش می‌شود که با کمک تعیین هایپرپارامترها این مشکل تا حدی حل می‌شود. در اینجا کمترین خلوصی که باید به آن برسیم تا الگوریتم متوقف شود، کمترین تعداد نمونه در هر نوع برای اعمال کلاس‌بندی روی اون نود و بیشترین تعداد برگ به عنوان هایپرپارامترها و توسط گرید سرج مقداردهی می‌شوند.

با آموزش مدل مقادیر زیر به عنوان مقادیر دقت و خطأ بدست امده اند.

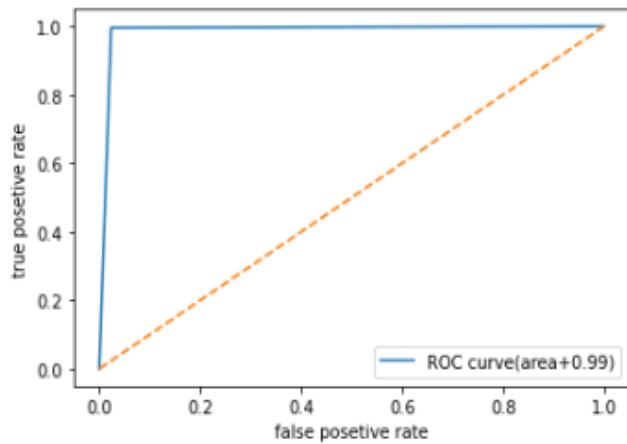
train score= 0.99, test score= 0.98

accuracy= 0.99, recall=1.00, precision= 0.98

f1= 0.99, MSE=0.01



شکل 17 ماتریس درهم ریختگی



شکل 18 ROC Curve

تنها ۱۲ عکس فیک را به اشتباه واقعی تشخیص داده و دو عکس واقعی رو عکس فیک تشخیص داده است که عملکردی عالی است همچنین مساحت زیر نمودار ROC بسیار نزدیک یک است که به معنی عملکرد عالی مدل می باشد .

رگرسیون لجستیک:

این الگوریتم یک خروجی احتمالی به ما می هد، سریع عمل می کند و برای داده هایی که به شکل خطی جدایزیرند بسیار خوب عمل می کند. اگر تعداد ویژگی ها زیادتر از نمونه ها باشد روش مناسبی نیست.

هایپرپارامتر ها:

C

معکوس رگولاریزیشن است و هر چقدر کوچکتر باشد الگوریتم رگولاریزیشن بهتری دارد
Penalty

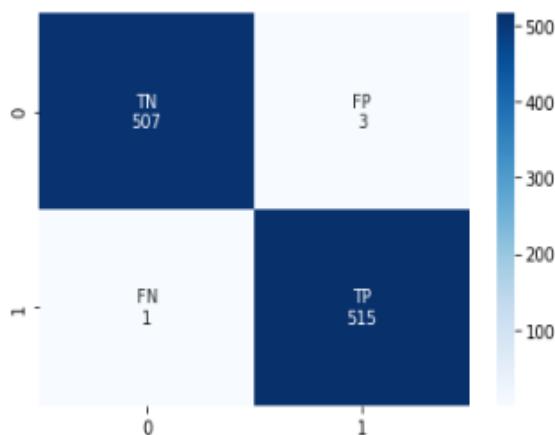
نوع تابع هزینه را مشخص می کند که هدف اصلی ان کنترل پیچیدگی مدل و جلوگیری از بیش برازش است.

با اموزش مدل مقادیر زیر به عنوان مقادیر دقیق و خطأ بدست آمده اند.

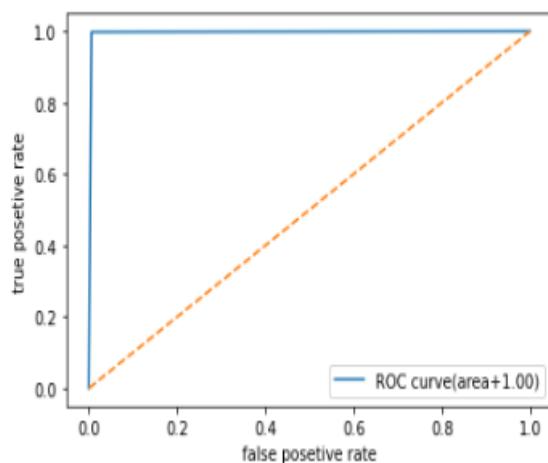
```
train score= 0.99, test score= 0.99
```

```
accuracy= 1.00, recall=1.00, precision= 0.99
```

f1= 1.00, MSE=0.00



شکل 19 ماتریس درهم ریختگی



ROC Curve 20

عملکرد ان نسبت به مدل قبلی بهتر است و تنها ۳ عکس فیک رو به عنوان واقعی تشخیص داده است و یک عکس واقعی را به اشتباه فیک تشخیص داده.

SVM

یک الگوریتم است که کلاس بندی را بر اساس محدوده‌ی نمونه‌ها انجام می‌دهد عملکرد خوبی برای داده‌هایی که به صورت خطی جدابذیرند دارد اگر کلاس‌ها همپوشانی زیادی با یکدیگر داشته باشند خوب عمل نمی‌کند که البته با انتخاب یک هسته و بردن داده‌ها در یک فضایی با ابعاد بزرگ‌تر می‌تواند این مشکل را حل کند. این الگوریتم از لحاظ محاسباتی پر هزینه است و بسیار وابسته به نوع انتخاب هسته است.

هایپرپارامتر ها:

C

یک پارامتر است که میزان تعمیم پذیری را تعیین می کند. میزان بزرگ تر آن به معنی این است که مدل بیشتر به داده های آموزشی وابسته می شود و سعی می کند تمام نقلات را به درستی دسته بندی کند. از طرفی دیگر مقدار کوچکتر ان سعی می کند مقدار حاشیه بزرگتری بین دسته ها ایجاد کند.

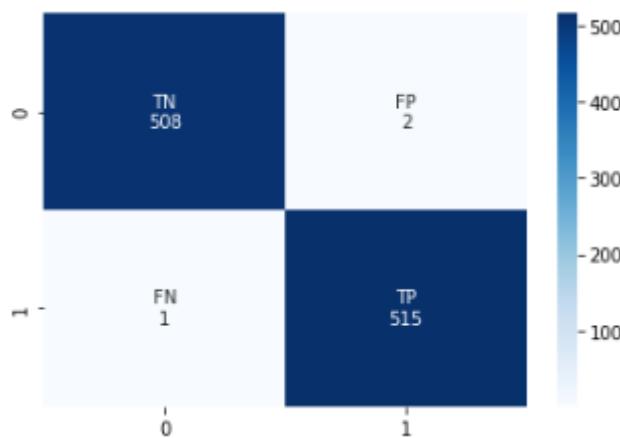
Kernel

نحوه ای تبدیل داده ها در فضای بالاتر را مشخص می کند که می تواند یکتابع خطی، چند جمله ای، شعاعی و ... باشد. استفاده از کرنل مناسب به دقت عملکرد مدل کمک می کند. با اموزش مدل مقادیر زیر به عنوان مقادیر دقت و خطأ بدست امده اند.

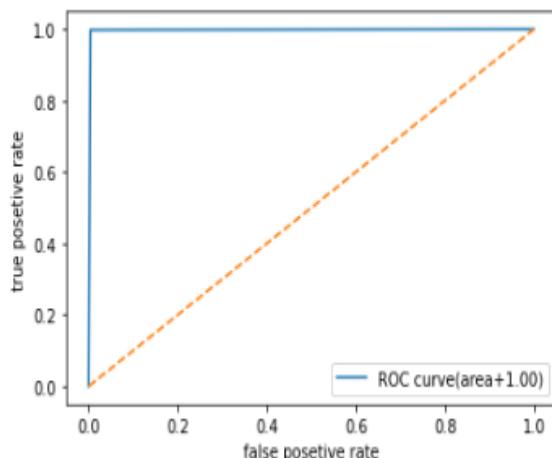
train score= 0.998, test score= 0.997

accuracy= 1.00, recall=1.00, precision= 1.00

f1= 1.00, MSE=0.00



شکل 21 ماتریس درهم ریختگی



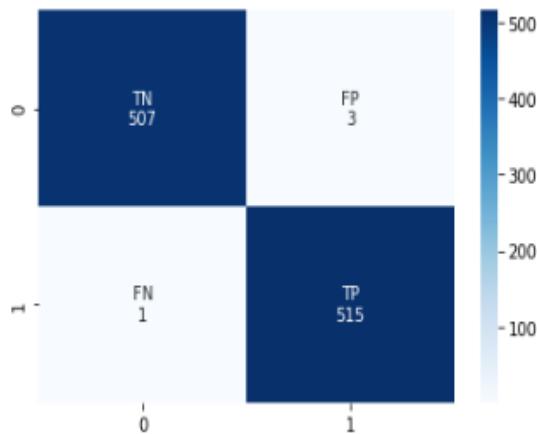
ROC Curve 22

Random Forest

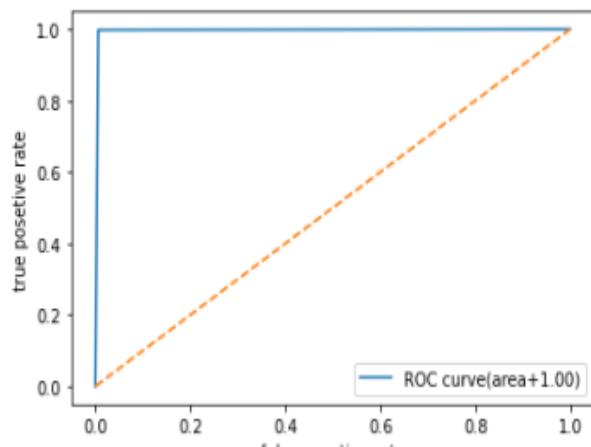
یک راه حل برای مشکل بیش برازش درخت تصمیم است. شامل تعداد زیادی درخت تصمیم می باشد که هر کدام ویژگی های خود را دارند و برای نتیجه ی نهایی بین نتایج درخت ها رای گیری می شود هایپر پارامترهایی که در اینجا تعیین شده اند شامل: تعداد درختان جنگل و روش بدست اوردن تعداد ویژگی در هر گام ساختن درخت می شوند.

با اموزش مدل مقادیر زیر به عنوان مقادیر دقت و خطأ بدست امده اند.

```
train score= 1.00, test score= 0.996  
accuracy= 1.00, recall=1.00, precision= 0.99  
f1= 1.00, MSE=0.00
```



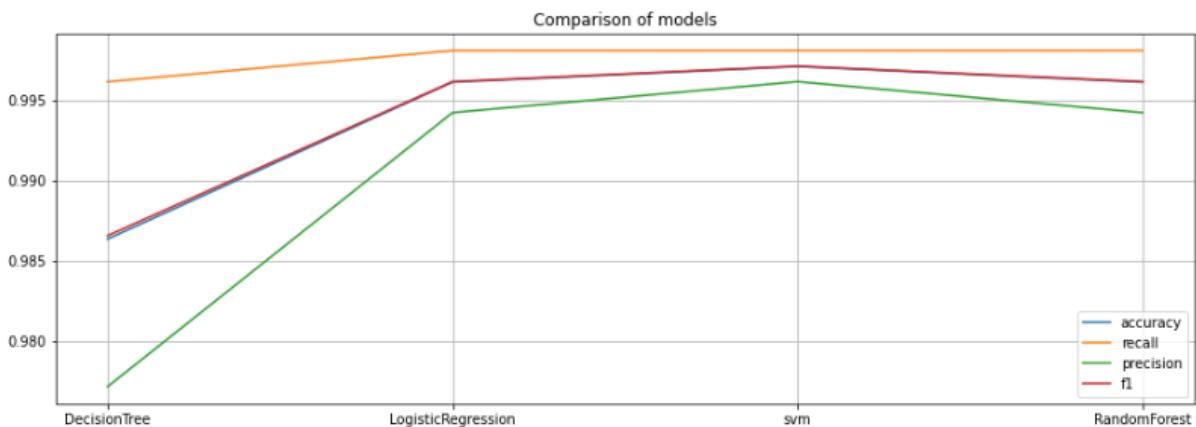
شکل 23 ماتریس درهم ریختگی



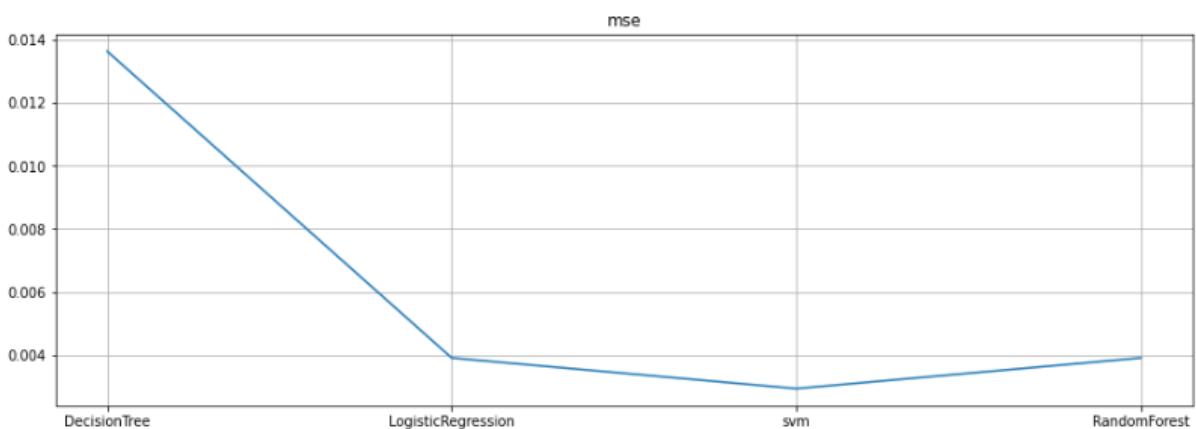
شکل 24 ROC Curve

این مدل تنها ۳ عکس فیک رو به عنوان واقعی تشخیص داده است و یک عکس واقعی را به اشتباه فیک تشخیص داده است.

در دو شکل پایین خطای دقت در تمامی مدل‌ها مقایسه شده که عملکرد درخت تصمیم از بقیه بدتر بوده



شکل 24 مقایسه مقادیر دقت در مدل‌ها



شکل 25 مقایسه مقادیر خطای دقت در مدل‌ها

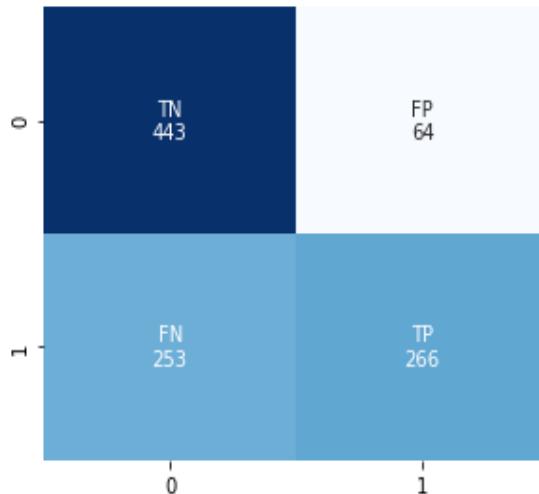
۲.۲-اموزش مدل برای ویژگی های استخراج شده

درخت تصمیم:

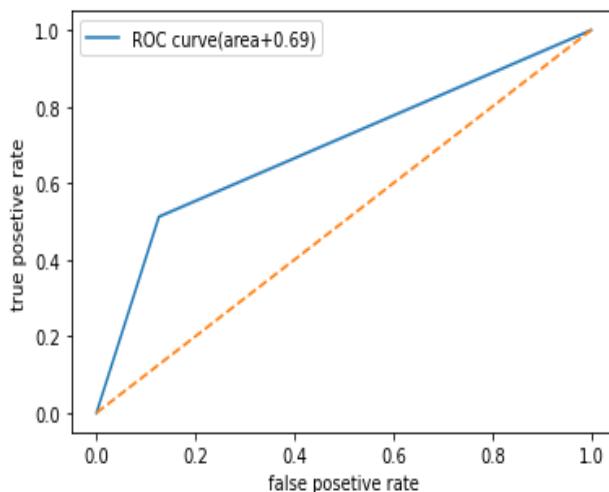
train score= 0.721, test score= 0.69

accuracy= 0.69, recall=0.51, precision= 0.81

f1= 0.63, MSE=0.31



شکل 26 ماتریس درهم ریختگی



ROC Curve 27

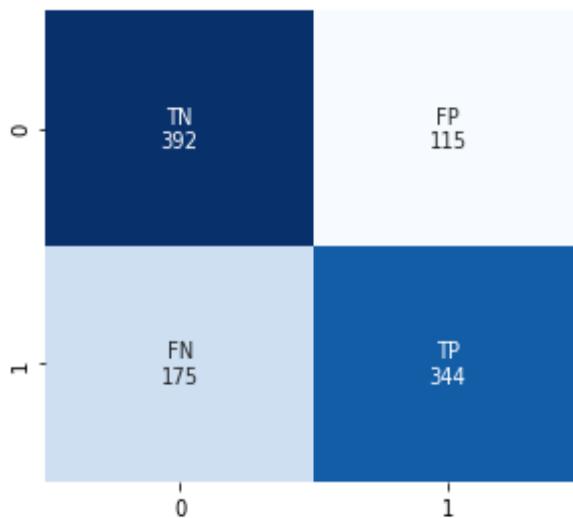
مساحت زیر نمودار ROC برابر ۰,۶۹ شده که نشانه ای از کم برازش است. در بالا بیشترین دقت مربوط به precision است با مقدار ۸۱ درصد که در واقع نشان می دهد عملکرد به گونه ای بوده که بیشتر عکس ها رو فیک تشخیص داده که این باعث پایین آمدن recall نیز شده است.

رگرسیون لجستیک:

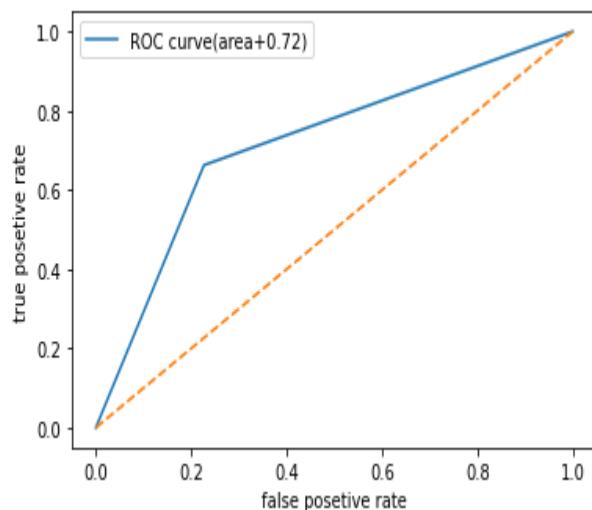
train score= 0.77, test score= 0.71

accuracy= 0.72, recall=0.66, precision= 0.75

f1= 0.70, MSE=0.28



شکل 28 ماتریس درهم ریختگی



ROC Curve 28

مساحت زیر نمودار ۰,۷۲ و کمی بیشتر از مدل قبل شده همانطور که گفتیم این نشانه‌ی کم برآش است و همچنین نشان می‌دهد که داه‌های ما به صورت خطی جداپذیر نیستند. ۱۱۵ عکس به اشتباه واقعی تشخیص داده و ۱۷۵ عکس واقعی را به عنوان عکس فیک تشخیص داده است با توجه به این

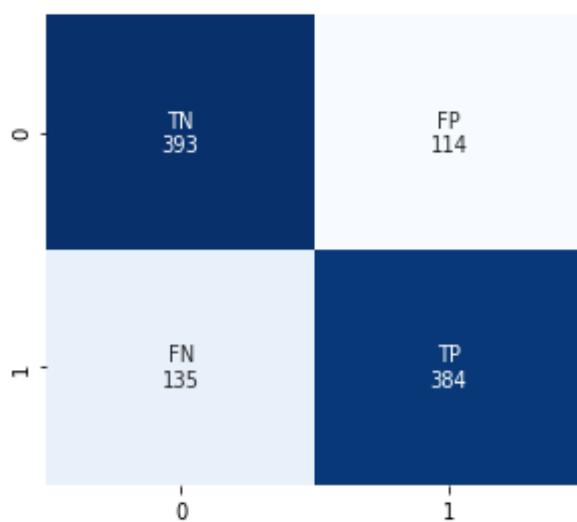
که در بالا هم precision بیشترین مقدار را داشت می توان گفت تشخیص عکس های واقعی نسبت به تشخیص عکس های فیک برای مدل سخت تر است.

SVM

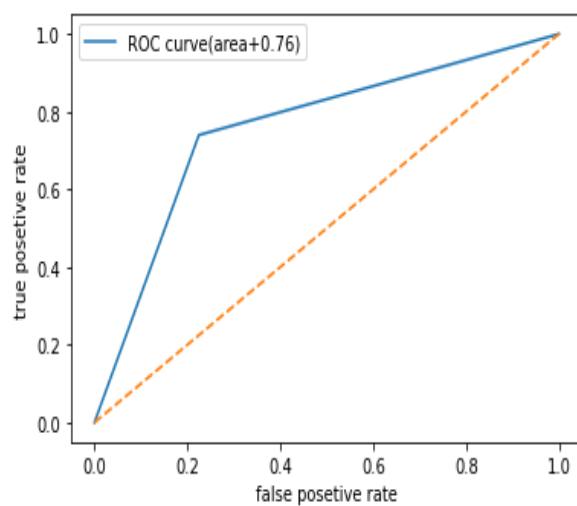
train score= 0.93, test score= 0.75

accuracy= 0.76, recall=0.74, precision= 0.77

f1= 0.76, MSE=0.24



شکل 29 ماتریس درهم ریختگی

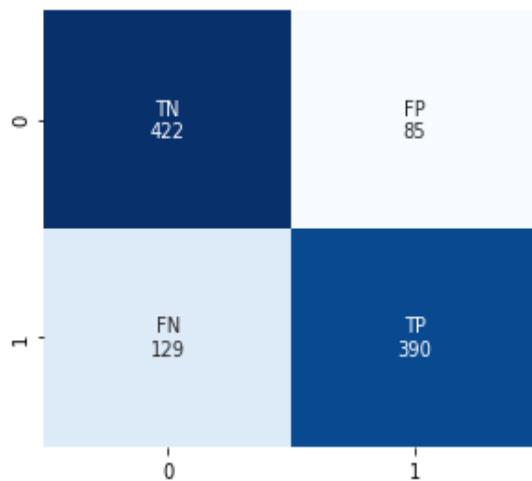


ROC Curve 30

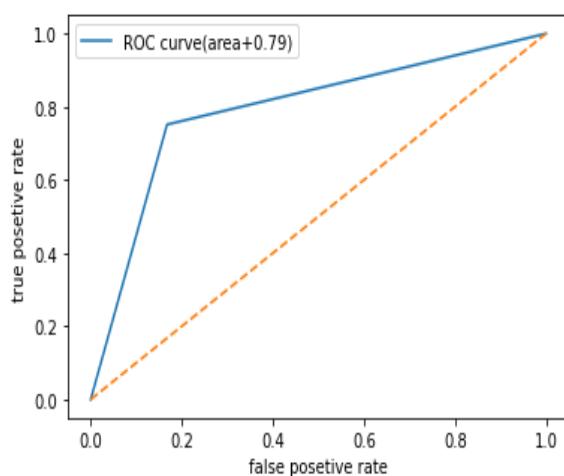
مساحت زیر نمودار ROC از مدل های قبلی بیشتر شده و از خط رندوم فاصله گرفته که خوب طبق انتظار بود چون این مدل با توجه به کرنل های خود قدرت بیشتری دارد. ولی چون دقت ترین و تست فاصله‌ی زیادی باهم دارند نشانه‌ی از بیش برازش است. در اینجا همانند دو مدل قبلی تعداد عکس‌های واقعی که به عنوان فیک تشخیص داده شده اند بیشتر از عکس‌های فیکی است که واقعی تشخیص داده شده‌اند.

:Random Forest

```
train score= 1.00, test score= 0.79
accuracy= 0.75, recall=0.75, precision= 0.82
f1= 0.78, MSE=0.21
```



شکل 31 ماتریس درهم ریختگی



ROC Curve 32

مقدار دقت روی ترین یک شده در حالی که دقت تست ۷۹ درصد است و این دو فاصله‌ی زیادی از هم دارند و نشانه بیش برآش است با این حال مساحت زیر نمودار ROC در بهترین حالت خود نسبت به مدل‌های قبلی است و برابر $0,79$ می‌باشد همچنین مقدار اشتباهات نسبت به مدل‌های قبلی کمتر است 85 عکس فیک به عنوان عکس واقعی و 129 عکس واقعی به عنوان عکس فیک درنظر گرفته شده‌اند.

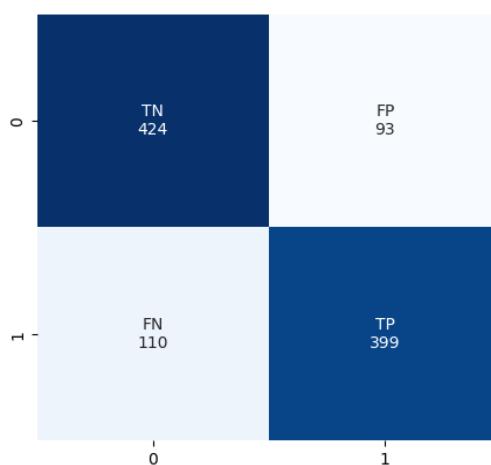
شبکه عصبی:

شبکه عصبی یک مدل محاسباتی است که از ساختار و عملکرد شبکه‌های عصبی بیولوژیکی مانند مغز انسان الهام گرفته شده است. یک شبکه عصبی از گره‌های به هم پیوسته‌ای به نام نورون‌ها تشکیل شده است که در لایه‌ها سازماندهی شده‌اند. در شبکه‌های عصبی از الگوریتم بهینه‌سازی و توابع فعال سازی استفاده می‌شود. توابع فعال سازی خروجی یک نورون را بر اساس ورودی آن تعیین می‌کنند. توابع فعال سازی رایج عبارتند از: سیگموئید، ReLU و tanh. انتخاب تابع فعال سازی می‌تواند بر توانایی شبکه در مدل‌سازی روابط غیرخطی و جریان گرادیان در طول آموزش تأثیر بگذارد. الگوریتم بهینه‌سازی تعیین می‌کند که چگونه شبکه وزن‌های خود را بر اساس گرادیان‌های محاسبه شده به روز می‌کند. الگوریتم‌های بهینه‌سازی محبوب عبارتند از نزول گرادیان تصادفی (SGD)، Adam و RMSprop. هر الگوریتم خواص و رفتارهای همگرایی متفاوتی دارد.

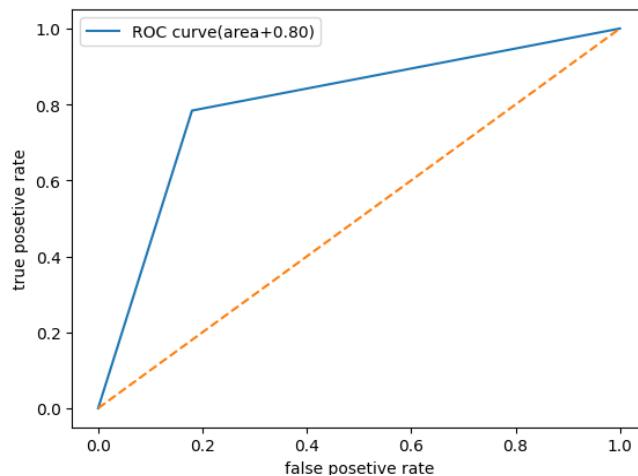
train score= 0.85, accuracy= 0.80,

recall=0.78, precision= 0.81

f1= 0.80, MSE=0.20



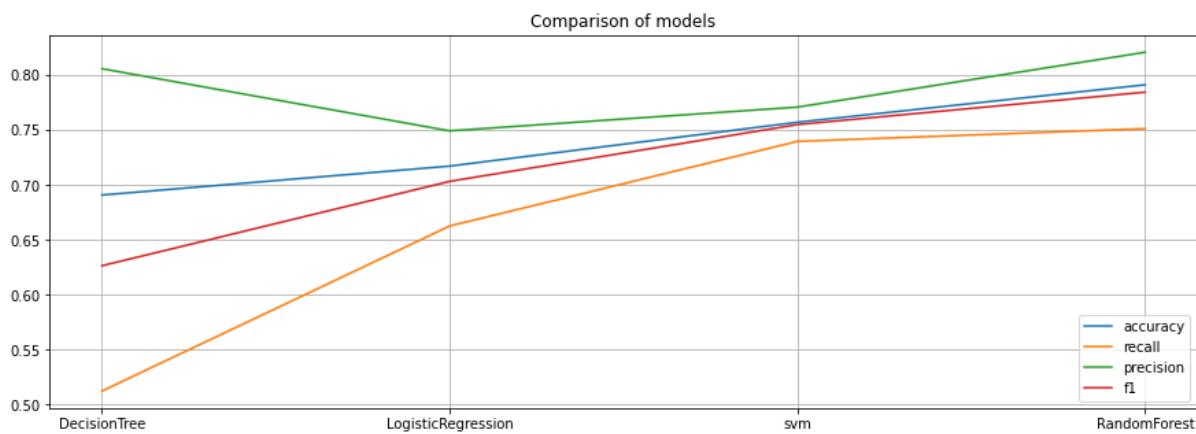
شکل 33 ماتریس درهم ریختگی



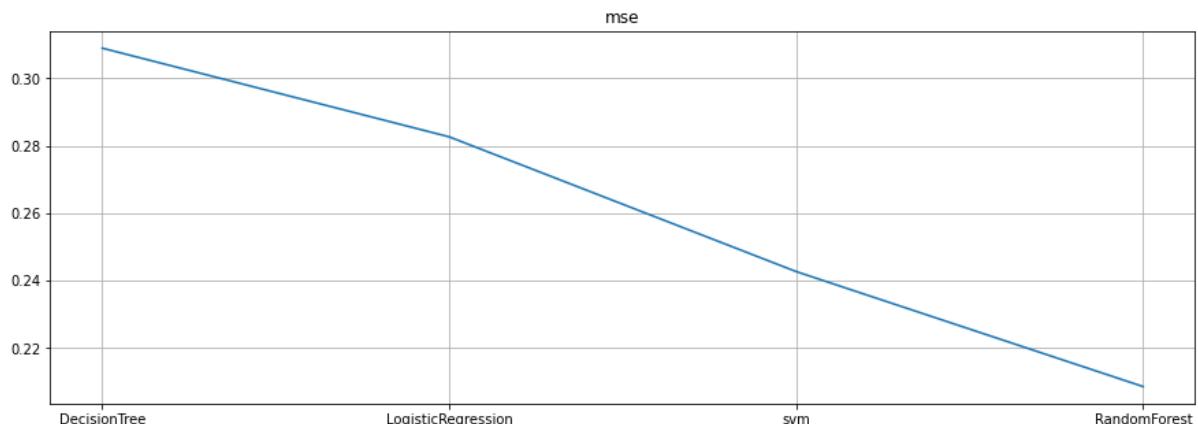
ROC Curve 34 شکل

در شبکه عصبی مساحت زیر نمودار ROC بیشتر از بقیه‌ی مدل‌هاست همچنین مطابق انتظار تمامی دقت‌ها نیز بیشتر می‌باشد.

در دو شکل پایین خطا و دقت در تمامی مدل‌ها مقایسه شده است.



شکل 34 مقایسه مقادیر دقت در مدل‌ها



شکل 35 مقایسه مقادیر خطأ در مدل ها

بخش 3 - خوشبندی

3.1 مدل خوشبندی برای ویژگی‌های داده شده

:GMM

خوشبندی GMM (مدل مخلوط گاووسی) که به عنوان خوشبندی مخلوط گاووسی نیز شناخته می‌شود، یک الگوریتم خوشبندی مبتنی بر مدل احتمالی است. برخلاف K-means یا خوشبندی سلسله مراتبی، که هر نقطه داده را به یک خوشبندی اختصاص می‌دهد، خوشبندی GMM اجازه می‌دهد تا تخصیص نرم صورت بگیرد، به این معنی که هر نقطه داده می‌تواند احتمال تعلق به خوشبندی‌های متعدد داشته باشد.

الگوریتم خوشبندی GMM فرض می‌کند که نقاط داده در هر خوشبندی از یک توزیع گاووسی پیروی می‌کنند. هدف ان برآورد پارامترهای این توزیع‌ها یعنی مانند میانگین و کوواریانس برای شناسایی خوشبندی‌های داده‌ها است. الگوریتم به طور متوالی احتمال مشاهده شده را با توجه به پارامترهای تخمین زده شده به حداقل می‌رساند.

در اینجا یک مرور کلی از الگوریتم خوشبندی GMM ذکر می‌شود:

1. الگوریتم را با انتخاب تصادفی تعداد خوشبندی‌ها و پارامترهای اولیه آنها مانند میانگین و کوواریانس مقداردهی اولیه کنید.

2. E-step (گام انتظار): احتمال تعلق هر نقطه داده به هر خوشبندی از برآورد فعلی پارامترها محاسبه کنید. این مرحله به هر نقطه داده یک احتمال تعلق به هر خوشبندی می‌دهد.

3. M-step (مرحله حداقل‌سازی): پارامترهای هر خوشبندی را با به حداقل رساندن احتمال محاسبات در مرحله 2 به روز کنید. این مرحله شامل برآورد مجدد میانگین‌ها، کوواریانس‌ها و prior های توزیع گاووسی است.

4. مراحل 2 و 3 را تا زمان همگرایی تکرار کنید (زمانی که برآورد پارامتر ثابت می‌شود یا معیار توقف از پیش تعريف شده برآورده می‌شود).

هنگامی که الگوریتم GMM همگرا می‌شود، به هر نقطه داده احتمال تعلق به هر خوشبندی بر اساس پارامترهای توزیع گاووسی اختصاص داده می‌شود. تخصیص سخت را می‌توان با انتخاب خوشبندی با بیشترین احتمال برای هر نقطه داده به دست آورد.

به طور کلی، خوشه بندی GMM یک الگوریتم قدرتمند برای کشف الگوهای خوبه های پنهان در داده ها است، به ویژه هنگام برخورد با توزیع های پیچیده و خوشه های همپوشان.

مزایای خوشه بندی GMM

1. انعطاف پذیری در شکل خوشه: خوشه بندی GMM می تواند خوشه ها را با اشکال دلخواه ضبط کند، زیرا نقاط داده را با استفاده از توزیع های گاوی مدل می کند. این باعث می شود که این روش برای مجموعه داده ها با ساختارهای خوشه ای پیچیده و غیر کروی مناسب باشد.
2. تخصیص نرم: خوشه بندی GMM تخصیص نرم را فراهم می کند، به این معنی که هر نقطه داده می تواند متعلق به خوشه های متعدد با احتمالات مختلف باشد. این انعطاف پذیری اجازه می دهد تا تخصیصهای خوشه ای ظریف تر باشند و زمانی که نقاط داده در نزدیکی مرزهای خوشه های مختلف قرار دارند می توانند مفید باشند.
3. براورد احتمال: خوشه بندی GMM نه تنها خوشه ها را شناسایی می کند بلکه احتمال عضویت هر نقطه داده در هر خوشه را نیز براورد می کند. این می تواند در برنامه هایی که در ان اطلاعات احتمالی مورد نیاز است، مانند تشخیص ناهنجاری یا میزان عدم اطمینان، ارزشمند باشد.
4. قدرتمندی نسبت به داده های پرت: خوشه بندی GMM نسبت به سایر الگوریتم های خوشه بندی نسبتاً قوی است. داده های پرت کمتر احتمال دارد که بر براورد مدل از توزیع ها تسلط داشته باشند و در نتیجه نتایج خوشه بندی قابل اعتماد تر شود.
5. توانایی اداره خوشه های نابرابر: خوشه بندی GMM می تواند خوشه هایی با اندازه های مختلف را اداره کند و فرض نمی کند که خوشه ها دارای واریانس برابر یا تعداد یکسانی از نقاط داده باشند.

معایب خوشه بندی GMM

1. حساسیت به مقداردهی اولیه: خوشه بندی GMM به پارامترهای اولیه حساس است و بسته به حدس اولیه می تواند به جواب های مختلف همگرا شود. ممکن است برای پیدا کردن یک بهینه جهانی به چندین بار مقداردهی اولیه نیاز باشد که می تواند زمان محاسباتی را افزایش دهد.
2. پیچیدگی محاسباتی: خوشه بندی GMM می تواند از نظر محاسباتی گران باشد، به ویژه برای مجموعه داده های بزرگ. این الگوریتم شامل براورد پارامترهای چندین توزیع گاوی است که نیاز به روش های بهینه سازی تکراری مانند الگوریتم Expectation-Maximization دارد.
3. مشکل با داده های با بعد بالا: خوشه بندی GMM می تواند در مواجهه با داده های با بعد بالا به دلیل "نفرین بعد" دچار چالشهایی شود. همانطور که تعداد بعد افزایش می یابد، براورد ماتریس های کوواریانس چالش برانگیزتر می شود و الگوریتم ممکن است برای پیدا کردن خوشه های معنی دار دچار مشکل شود.

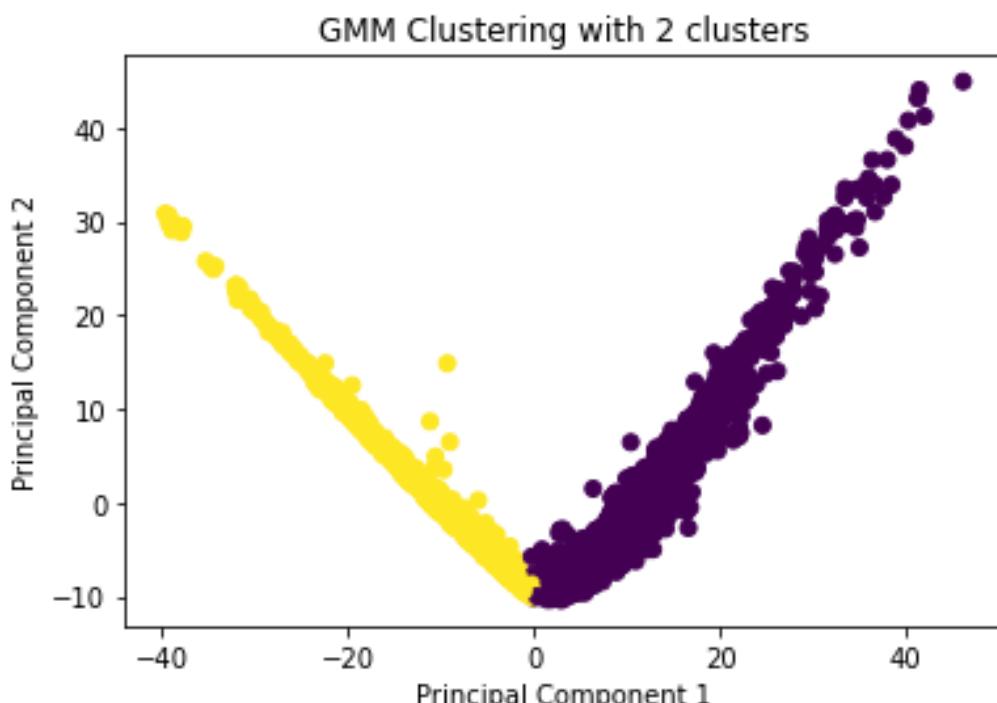
4. تعیین دستی اعداد خوشه ای: بر خلاف برخی از الگوریتم های خوشه بندی دیگر، خوشه بندی GMM نیاز به دانش قبلی یا تعیین دستی تعداد خوشه ها دارد. انتخاب تعداد مناسب خوشه ها می تواند ذهنی باشد و ممکن است نیاز به تخصص دامنه یا تکنیک های اعتبار سنجی اضافی داشته باشد.

5. تفسیر پذیری: خوشه بندی GMM یک ساختار سلسله مراتبی ساده مانند خوشه بندی سلسله مراتبی ارائه نمی دهد. تفسیر نتایج خوشه بندی GMM می تواند کمتر شهودی باشد، زیرا خوشه ها به جای دندروگرام با توزیع گاووسی نشان داده می شوند.

با ابتدا کتابخانه ها و داده های مربوط را وارد می کنیم؛ سپس به ترتیب اقدام به فیت کردن GMM برای ۳، ۶، ۹ و ۵۰ کلاستر می کنیم. با کمک PCA ابعاد را کاهش داده و نمودار کلاسترها را در دو بعد رسم می کنیم سپس برای هر بخش و برای هر کلاستر سه عدد (شماره سطر داده) به صورت تصادفی استخراج کرده و بدین ترتیب برای هر خوشه به سه تصویر تصادفی از آن خوشه می رسمیم. نتیجه را در جداول زیر نمایش می دهیم.

به طور کلی، ضریب Silhouette از مقدار 0.29 برای دو خوشه تا مقدار 0.13 برای 50 خوشه کاهش می یابد.

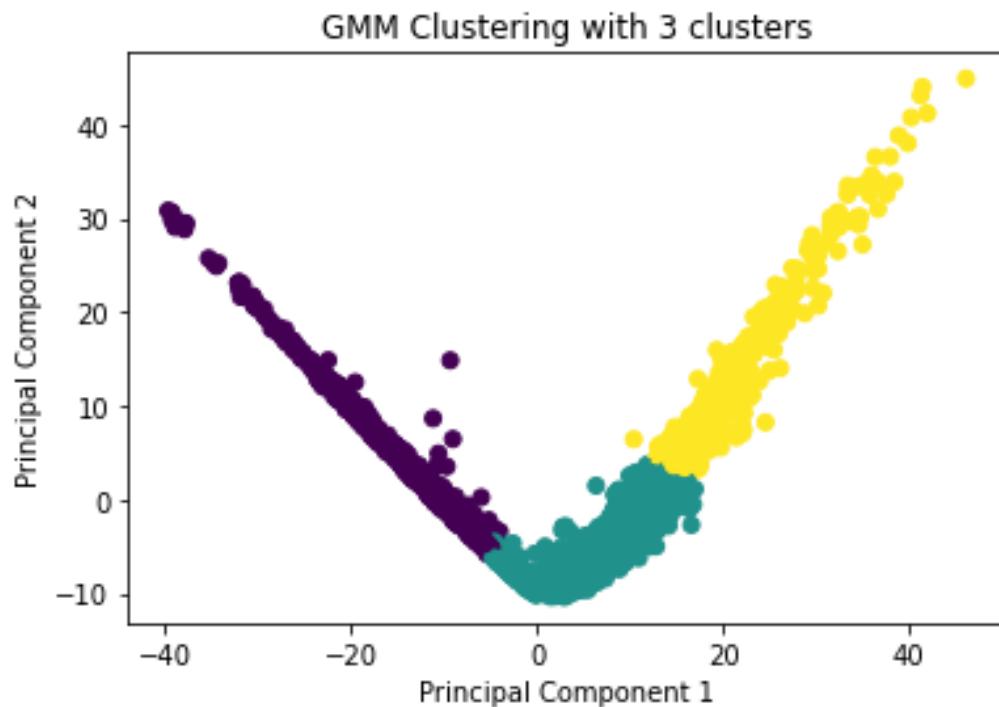
• نتایج دو خوشه



نقاط تصادفی به صورت زیر است

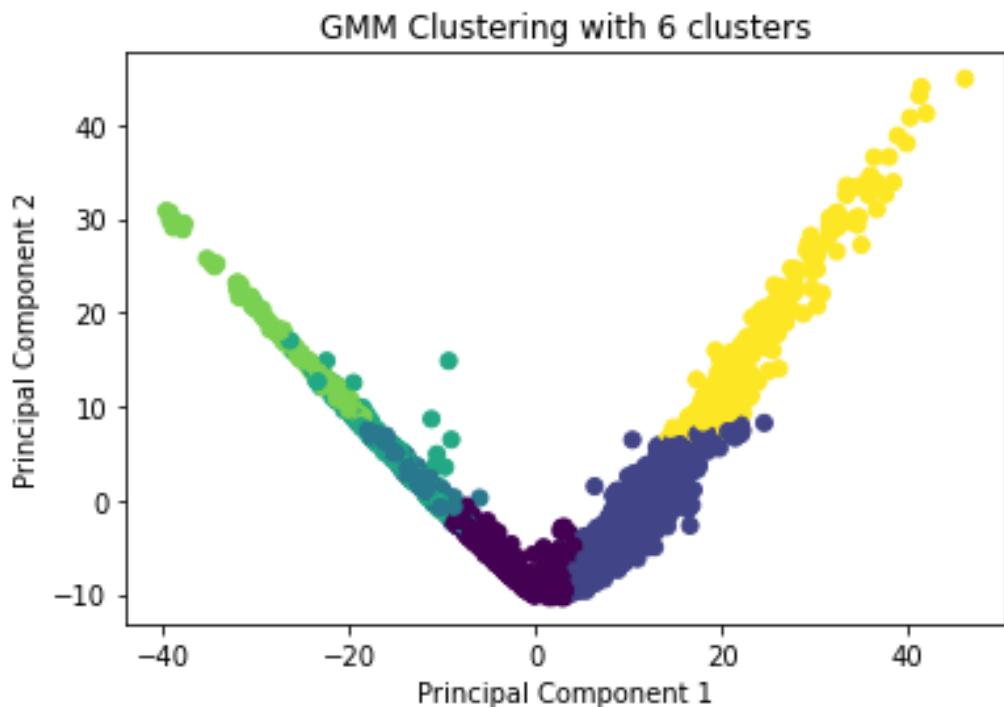
cluster	Image 1	Image 2	Image 3
0			
label	Fake sea	Fake mountain	Fake forest
1			
label	Real forest	Fake sea	Real sea

Random points from GMM cluster 0: [1022, 2819, 2229]
 Random points from GMM cluster 1: [2648, 1365, 1756]



```
Random points from GMM cluster 0: [1683, 606, 708]  
Random points from GMM cluster 1: [2250, 2388, 3301]  
Random points from GMM cluster 2: [712, 3159, 2835]
```

cluster	Image 1	Image 2	Image 3
0			
Label	Fake sea	Fake sea	Real maontain
1			
Label	Real mountain	Real sea	Fake sea
2			
label	Fake forest	Real mountain	Fake sea

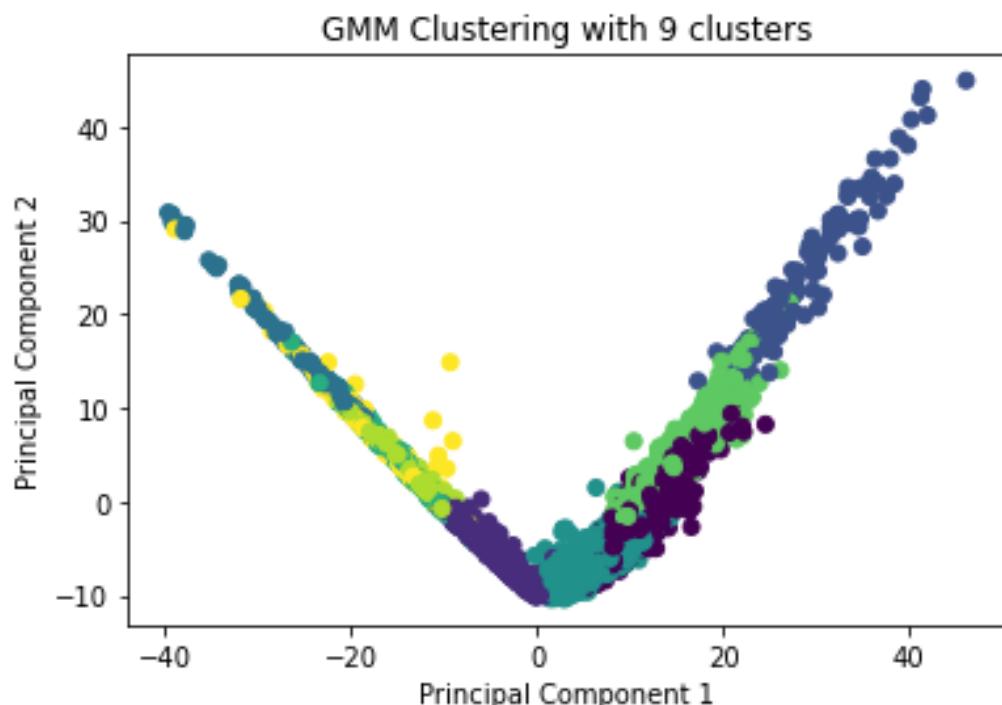


```
Random points from GMM cluster 0: [2576, 994, 526]  
Random points from GMM cluster 1: [1643, 2148, 2366]  
Random points from GMM cluster 2: [2001, 2806, 2623]  
Random points from GMM cluster 3: [2173, 3002, 2208]  
Random points from GMM cluster 4: [2616, 89, 2458]  
Random points from GMM cluster 5: [2470, 1344, 674]
```

cluster	Image 1	Image 2	Image 3
0			
Label	Real sea	Fake forest	Fake sea
1			
Label	Fake forest	Fake forest	fake sea
2			
label	real sea	Real forest	Real mountain
3			
label	Real mountain	Real forest	Fake sea
4			
Label	fake forest	Real forest	Real forest

5				
Label	Real mountain	Fake mountain	Real forest	

نتائج نه خوش
•



```

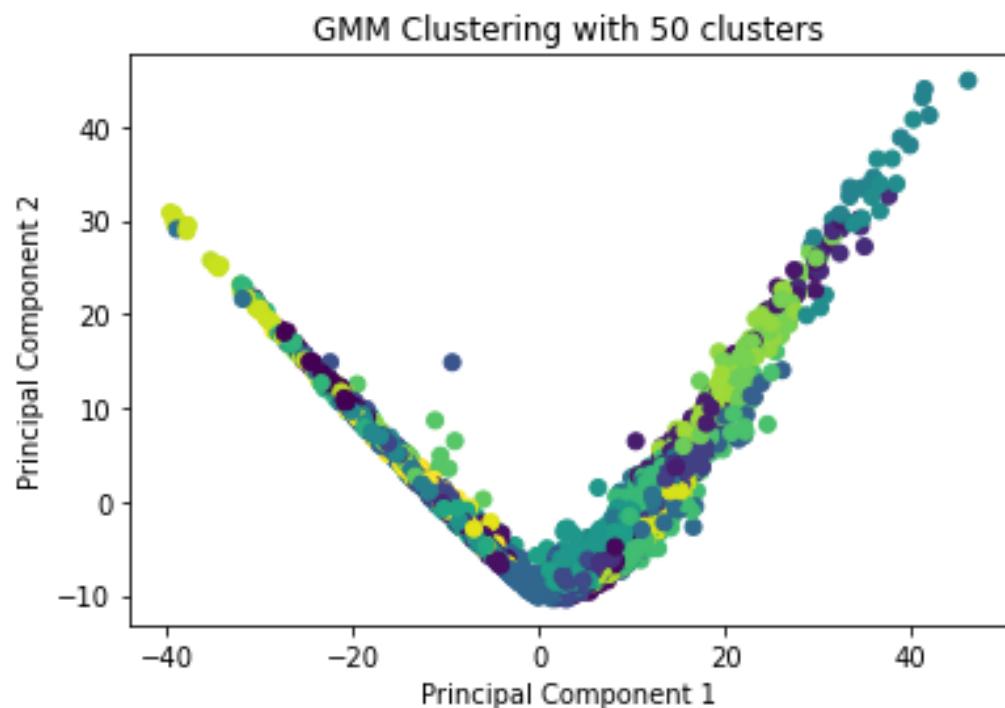
Random points from GMM cluster 0: [416, 164, 2619]
Random points from GMM cluster 1: [3054, 3384, 2223]
Random points from GMM cluster 2: [6, 1097, 290]
Random points from GMM cluster 3: [36, 248, 758]
Random points from GMM cluster 4: [553, 2007, 535]
Random points from GMM cluster 5: [2560, 2512, 2342]
Random points from GMM cluster 6: [3029, 444, 2894]
Random points from GMM cluster 7: [1588, 2149, 3326]
Random points from GMM cluster 8: [3086, 262, 175]

```

cluster	Image 1	Image 2	Image 3
---------	---------	---------	---------

0	Fake mountain	Real mountain	Real mountain
1	Real mountain	Fake forest	Real forest
2	Real forest	Fake forest	Real sea
3	Fake forest	Real mountain	Real sea
4	Fake sea	Fake mountain	Fake forest
5	Fake sea	Real sea	Fake mountain
6	Real sea	Real mountain	Real mountain
7	Real sea	Real forest	Fake forest
8	Fake forest	Fake mountain	Fake mountain

• نتایج پنجاه خوش



```

Random points from GMM cluster 0: [1546, 2157, 122]
Random points from GMM cluster 1: [3174, 3262, 1759]
Random points from GMM cluster 2: [1855, 1643, 238]
Random points from GMM cluster 3: [2587, 1058, 1427]
Random points from GMM cluster 4: [742, 65, 1298]
Random points from GMM cluster 5: [1063, 1780, 1994]
Random points from GMM cluster 6: [42, 681, 344]
Random points from GMM cluster 7: [1801, 1597, 1324]
Random points from GMM cluster 8: [2621, 2059, 1970]
Random points from GMM cluster 9: [2882, 748, 1841]
Random points from GMM cluster 10: [1142, 1595, 2514]
Random points from GMM cluster 11: [3126, 2890, 2196]
Random points from GMM cluster 12: [2874, 722, 1722]
Random points from GMM cluster 13: [994, 1280, 2187]
Random points from GMM cluster 14: [879, 2800, 3018]
Random points from GMM cluster 15: [824, 1737, 1439]
Random points from GMM cluster 16: [2234, 617, 3367]
Random points from GMM cluster 17: [3315, 1592, 2522]

```

Hierarchical

خوشه بندی سلسله مراتبی یک الگوریتم محبوب است که در یادگیری ماشین بدون ناظارت برای گروه بندی نقاط داده مشابه به خوشه ها بر اساس شباهت ها یا فاصله های جفتی انها استفاده می شود. این روش یک ساختار سلسله مراتبی از خوشه ها ایجاد می کند که می تواند با استفاده از دندروگرام ترسیم شود.

الگوریتم با در نظر گرفتن هر نقطه داده به عنوان یک خوشه فردی شروع می شود. سپس، به طور متوالی خوشه ها را بر اساس نزدیکی انها ادغام می کند تا زمانی که تمام نقاط داده متعلق به یک خوشه واحد شوند و یا یک معیار توقف براورده شود. دو نوع اصلی از الگوریتم های خوشه بندی سلسله مراتبی وجود دارد: "پایین به بالا" و "بالا به پایین".

خوشه بندی سلسله مراتبی پایین به بالا با هر نقطه داده به عنوان یک خوشه جداگانه شروع می شود و سپس دو خوشه نزدیک را به صورت متوالی ادغام می کند تا زمانی که تمام نقاط داده متعلق به یک خوشه واحد باشند. نزدیکی بین خوشه ها معمولاً با استفاده از یک متریک فاصله مانند فاصله اقلیدسی یا فاصله منهتن اندازه گیری می شود. معیارهای مختلف پیوند را می توان برای تعیین فاصله بین خوشه ها استفاده کرد، مانند پیوند تک (حداقل فاصله بین هر دو نقطه در خوشه ها)، پیوند کامل (حداکثر فاصله بین هر دو نقطه در خوشه ها) یا میانگین پیوند (فاصله متوسط بین تمام جفت نقاط در خوشه ها).

از سوی دیگر، خوشه بندی سلسله مراتبی بالا به پایین با تمام نقاط داده در یک خوشه واحد شروع می شود و خوشه ها را به خوشه های کوچکتر تقسیم می کند تا زمانی که هر خوشه تنها شامل یک نقطه داده باشد. این فرایند شامل انتخاب یک خوشه و تقسیم آن بر اساس برخی معیارها مانند به حداقل رساندن فاصله بین خوشه ای یا به حداقل رساندن واریانس درون خوشه ای است.

هر دو روش خوشه بندی سلسله مراتبی می تواند برای ساخت یک دندروگرام استفاده شود که نشان دهنده ساختار سلسله مراتبی خوشه ها است. دندروگرام به ما اجازه می دهد تا فرایند خوشه بندی را تجسم کنیم و در مورد تعداد مناسب خوشه ها تصمیم بگیریم.

خوشه بندی سلسله مراتبی دارای مزایای متعددی است، از جمله توانایی آن در نشان دادن ساختار سلسله مراتبی داده ها و انعطاف پذیری آن در اداره معیارهای مختلف فاصله و معیارهای پیوند. با این حال، می تواند از لحاظ محاسباتی برای مجموعه داده های بزرگ گران باشد و نتایج خوشه بندی ممکن است بسته به انتخاب متريک فاصله و معيار پیوند متفاوت باشد.

به طور کلی، خوشه بندی سلسله مراتبی یک الگوريتم مفید برای بررسی و درک الگوهای گروه بندی طبیعی در یک مجموعه داده است که بینش هایی را در مورد روابط و شباهت های نقاط داده ارائه می دهد.

مزایای خوشه بندی سلسله مراتبی:

1. ساختار سلسله مراتبی: یکی از مزایای اصلی خوشه بندی سلسله مراتبی این است که یک ساختار سلسله مراتبی یا دندروگرام تولید می کند که نشان دهنده روابط بین نقاط داده است. این ساختار می تواند بینش هایی را در مورد اساس سازمان و شباهت های درون داده ها ارائه دهد.
2. هیچ فرضیه ای در مورد تعداد خوشه ها وجود ندارد: خوشه بندی سلسله مراتبی بر خلاف برخی از الگوريتم های خوشه بندی دیگر نیازی به مشخص کردن تعداد خوشه ها از قبل ندارد. این روش به طور خودکار تعداد خوشه ها را بر اساس داده ها و روش پیوند انتخاب شده تعیین می کند.
3. انعطاف پذیری در روش های پیوند: خوشه بندی سلسله مراتبی روش های مختلف پیوند مانند پیوند تک، پیوند کامل و پیوند متوسط را ارائه می دهد. این انعطاف پذیری به کاربران اجازه می دهد تا روشی را انتخاب کنند که به بهترین وجه مناسب با داده های انها و خواص مورد نظر خوشه ها باشد.

4. تفسیرپذیری: دندروگرام تولید شده توسط خوشه بندی سلسله مراتبی یک نمایش بصری از فرایند خوشه بندی را فراهم می کند و تفسیر و درک نتایج را اسان تر می کند. این روش شناسایی خوشه در سطوح مختلف را برای کاربران ممکن می سازد.

معایب خوشه بندی سلسله مراتبی:

1. پیچیدگی محاسباتی: خوشه بندی سلسله مراتبی می تواند از نظر محاسباتی گران باشد، به ویژه برای مجموعه داده های بزرگ. زمان و حافظه مورد نیاز با تعداد نقاط داده افزایش می یابد و برای مجموعه داده های بسیار بزرگ کمتر عملی است.

2. حساسیت به نویز و پرتی: خوشه بندی سلسله مراتبی به نویز و پرتی داده حساس است که می تواند تاثیر قابل توجهی بر دندروگرام و تخصیص خوشه ای داشته باشد. داده های پرت می توانند اندازه گیری های شباهت را تحریف کنند و منجر به نتایج خوشه بندی کمتر از حد مطلوب شوند.

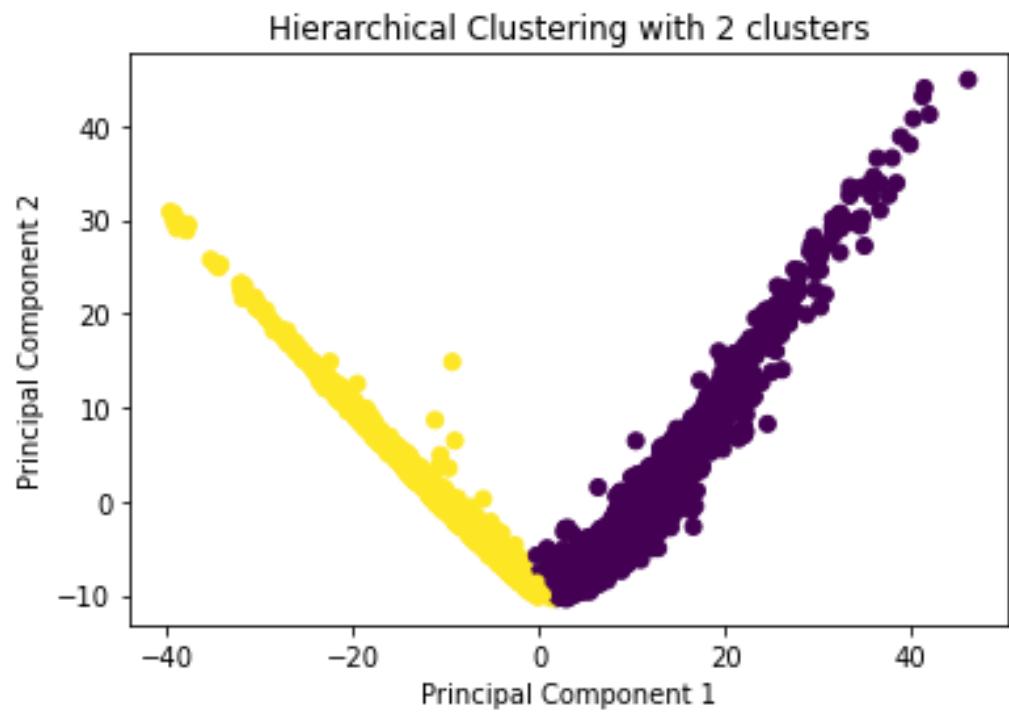
3. عدم مقیاس پذیری: همانطور که قبلا ذکر شد، پیچیدگی محاسباتی خوشه بندی سلسله مراتبی می تواند مقیاس پذیری ان را محدود کند. زمان و حافظه مورد نیاز الگوریتم ان را برای مجموعه داده های بزرگ و یا برنامه های کاربردی که در ان پردازش زمان واقعی لازم است، نامناسب می کند.

4. عدم انعطاف پذیری در شکل خوشه: خوشه بندی سلسله مراتبی تمایل به تولید خوشه های کروی دارد. اگر داده ها دارای اشکال خوشه ای غیر کروی یا پیچیده باشند، خوشه بندی سلسله مراتبی ممکن است قادر به گرفتن این ساختارها به طور موثر نباشد.

5. دشواری در اداره داده های با ابعاد بالا: خوشه بندی سلسله مراتبی می تواند با داده های بعدی بالا دچار مشکل شود زیرا محاسبات شباهت یا عدم شباهت در ابعاد بالاتر کمتر قابل اعتماد است. داده های با ابعاد بالا می توانند منجر به "نفرین ابعاد" شوند و منجر به نتایج خوشه بندی کمتر معنی داری شوند.

در این مدل نیز ضریب Silhouette از مقدار 0.29 برای دو خوشه تا مقدار 0.09 برای 50 خوشه کاهش می یابد.

• نتایج دو خوشه

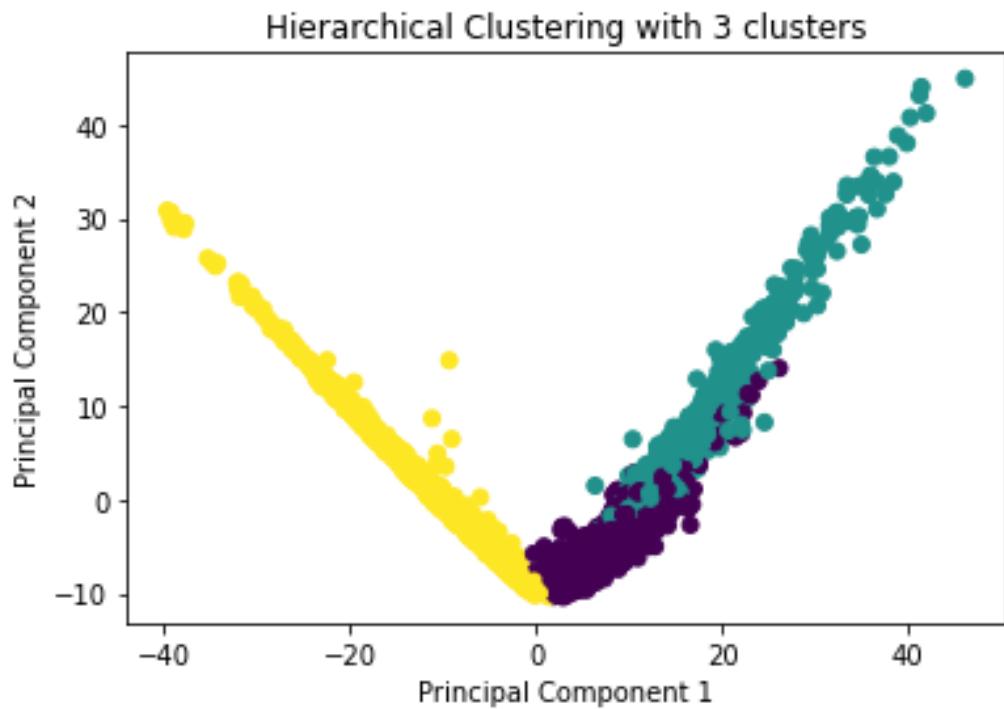


نقاط تصادفی به صورت زیر است

```
.... pit.show()
Random points from HAC cluster 0: [467, 625, 6]
Random points from HAC cluster 1: [499, 3290, 370]
```

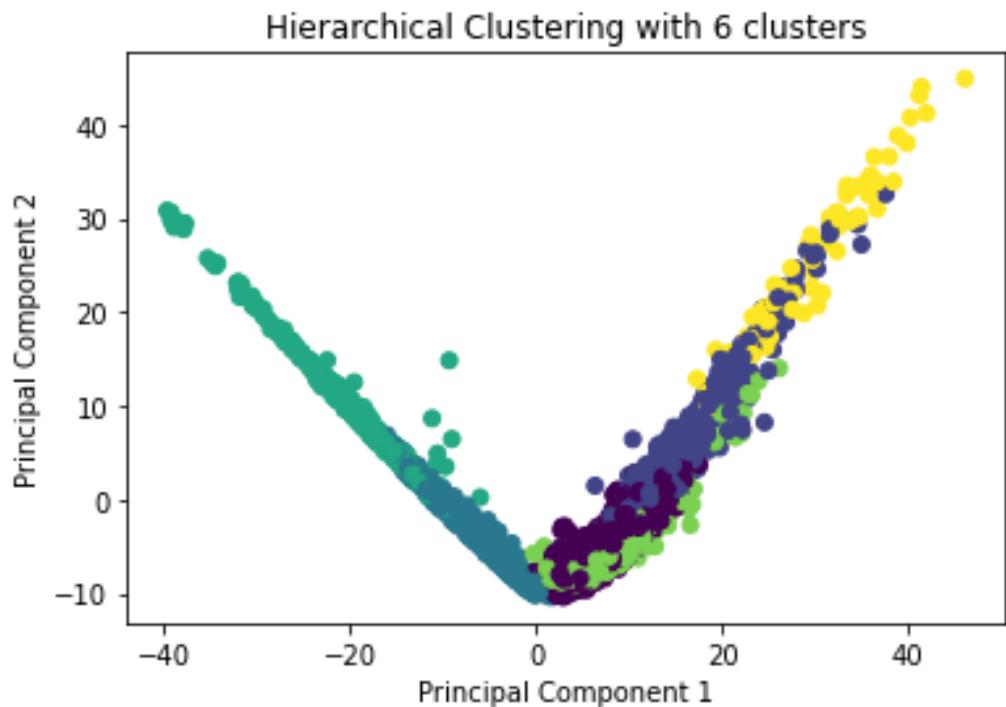
cluster	Image 1	Image 2	Image 3
0			
label	Fake mountain	Fake sea	Real forest
1			
label	Fake forest	Fake forest	Fake forest

نتایج سه خوشه •

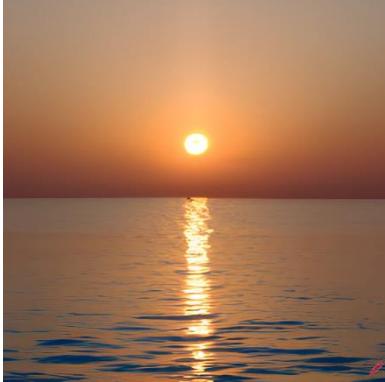
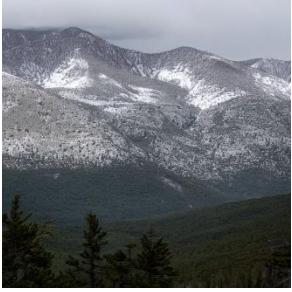


```
Random points from HAC cluster 0: [1593, 829, 1254]  
Random points from HAC cluster 1: [110, 2329, 2875]  
Random points from HAC cluster 2: [356, 57, 92]
```

cluster	Image 1	Image 2	Image 3
0			
Label	Fake sea	Fake forest	Real sea
1			
Label	Real mountain	Fake mountain	Real mountain
2			
label	Fake forest	Fake sea	Real sea

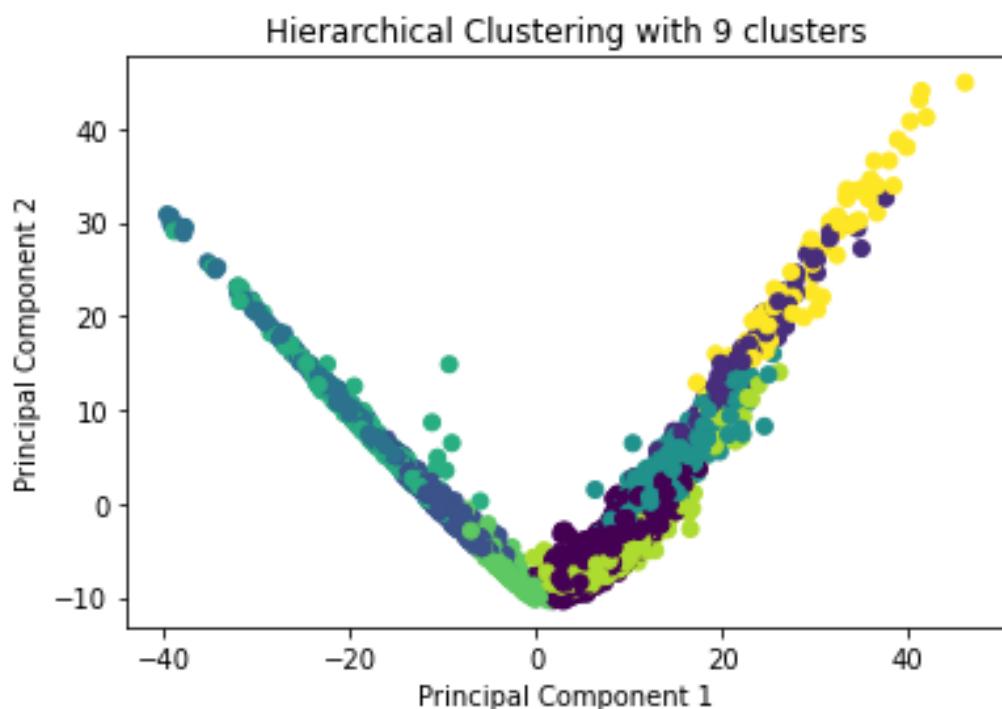


```
Random points from HAC cluster 0: [3153, 348, 2648]  
Random points from HAC cluster 1: [2302, 2366, 1071]  
Random points from HAC cluster 2: [1529, 1339, 2235]  
Random points from HAC cluster 3: [267, 2810, 794]  
Random points from HAC cluster 4: [2184, 813, 2847]  
Random points from HAC cluster 5: [1588, 1236, 755]
```

cluster	Image 1	Image 2	Image 3
0			
Label	fake sea	Fake mountain	Real forest
1			
Label	real forest	Fake sea	fake sea
2			
label	real forest	Fake mountain	fake forest
3			
label	Real forest	Real sea	Fake sea
4			
Label	Fake sea	Real mountain	Real forest

5			
Label	Real sea	Fake mountain	Fake mountain

نتایج نه خوشہ •

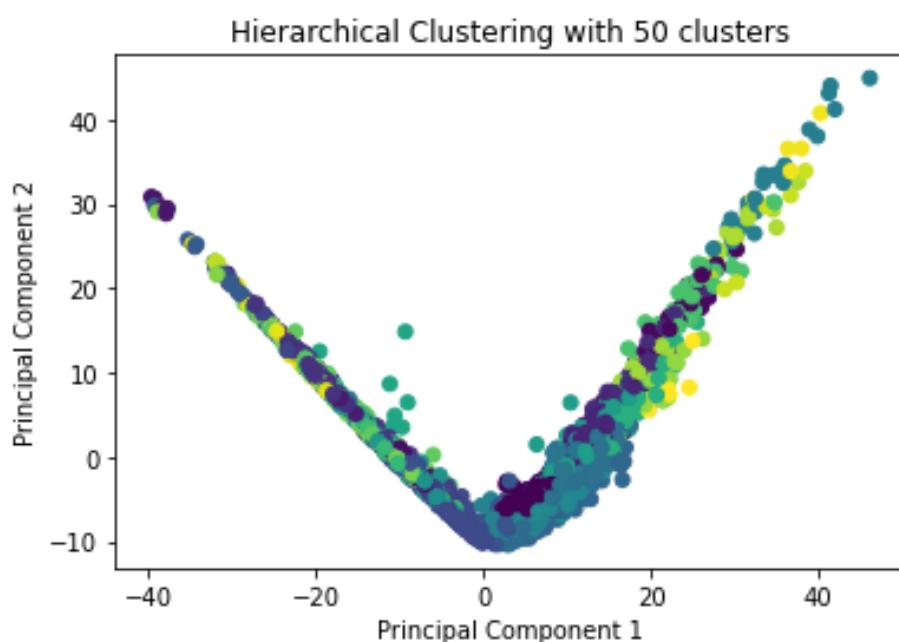


```
Random points from HAC cluster 0: [1262, 458, 3382]
Random points from HAC cluster 1: [106, 1450, 2383]
Random points from HAC cluster 2: [839, 2962, 3149]
Random points from HAC cluster 3: [895, 2540, 2810]
Random points from HAC cluster 4: [1614, 1342, 392]
Random points from HAC cluster 5: [2731, 356, 1567]
Random points from HAC cluster 6: [3403, 1806, 691]
Random points from HAC cluster 7: [376, 2088, 838]
Random points from HAC cluster 8: [552, 2133, 1904]
```

cluster	Image 1	Image 2	Image 3
0	Real mountain	Real forest	Fake mountain
1	Real mountain	Real mountain	Fake forest

2	Fake sea	Real sea	Fake mountain
3	Fake forest	Fake forest	Real sea
4	Fake mountain	Real mountain	Real sea
5	Real forest	Fake forest	Real mountain
6	Fake forest	Real forest	Fake mountain
7	Fake forest	Fake mountain	Real sea
8	Fake mountain	Fake sea	Real sea

نتایج پنجاه خوشہ •



```

Random points from HAC cluster 0: [832, 408, 2051]
Random points from HAC cluster 1: [1440, 1786, 904]
Random points from HAC cluster 2: [839, 3220, 1969]
Random points from HAC cluster 3: [2224, 3, 1809]
Random points from HAC cluster 4: [2332, 1266, 2862]
Random points from HAC cluster 5: [1947, 2400, 3297]
Random points from HAC cluster 6: [2087, 3362, 1766]
Random points from HAC cluster 7: [1633, 3346, 3267]
Random points from HAC cluster 8: [2620, 3164, 1093]
Random points from HAC cluster 9: [1575, 2368, 2465]
Random points from HAC cluster 10: [232, 462, 1013]
Random points from HAC cluster 11: [1965, 1813, 2869]
Random points from HAC cluster 12: [2470, 3329, 1235]
Random points from HAC cluster 13: [2144, 886, 184]
Random points from HAC cluster 14: [2117, 2254, 15]
Random points from HAC cluster 15: [130, 1476, 2517]
Random points from HAC cluster 16: [2100, 792, 3101]
Random points from HAC cluster 17: [3266, 3289, 2799]
Random points from HAC cluster 18: [1120, 3380, 943]
Random points from HAC cluster 19: [917, 41, 2258]
Random points from HAC cluster 20: [1834, 2241, 2536]

```

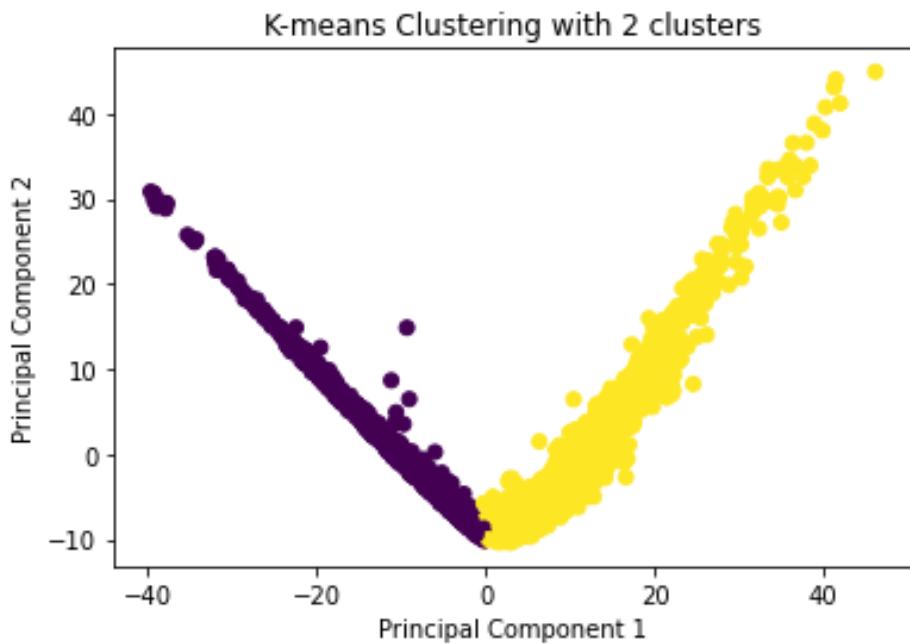
:k-means

K-means یک الگوریتم خوشبندی تکراری است که هدف آن تقسیم یک مجموعه داده به K خوشبندی مجزا است. الگوریتم با مقداردهی اولیه تصادفی K نقاط در فضای ویژگی، به نام centroids که مراکز خوشبندی را نشان می‌دهد، شروع می‌شود. سپس به طور مکرر دو مرحله را انجام می‌دهد: تخصیص و به روز رسانی. در مرحله تخصیص، هر نقطه داده بر اساس یک متریک فاصله، معمولاً فاصله اقلیدسی، به نزدیکترین مرکز تخصیص داده می‌شود. این مرحله خوشبندی K را ایجاد می‌کند که هر نقطه داده متعلق به خوشبندی مرکز با نزدیکترین مرکز است. در مرحله به روز رسانی، مرکزهای خوشبندی K با در نظر گرفتن میانگین تمام نقاط داده اختصاص داده شده به هر خوشبندی دوباره محاسبه می‌شوند. این مرکز جدید به مرکز به روز شده خوشبندی مرتبه خود تبدیل می‌شود.

الگوریتم به تکرار بین مراحل تخصیص و به روز رسانی تا زمان همگرایی ادامه می‌دهد. همگرایی معمولاً زمانی حاصل می‌شود که مرکزها دیگر تغییر قابل توجهی نداشته باشند یا زمانی که به حداقل تعداد تکرار رسیده باشند. هنگامی که الگوریتم همگرا شد، هر نقطه داده بر اساس مرکز نهایی به یک خوشبندی اختصاص داده می‌شود. خوشبندی حاصل، داده‌ها را به K گروههای مجزا جدا می‌کند، که در آن نقاط داده در هر خوشبندی بیشتر به یکدیگر شبیه هستند تا در سایر خوشبندی‌ها.

در این مدل نیز همانند gmm ضریب Silhouette از مقدار 0.29 برای دو خوشبندی تا مقدار 0.13 برای 50 خوشبندی کاهش می‌یابد.

• نتایج دو خوشبندی

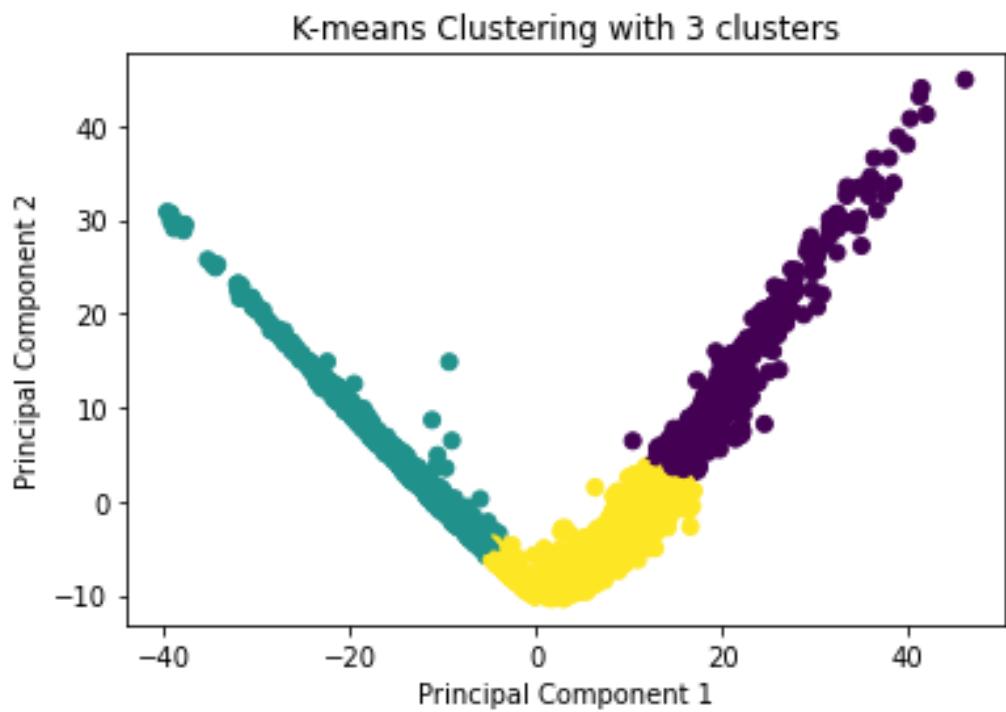


نقاط تصادفی به صورت زیر است

Random points from KMM cluster 0: [1807, 537, 3061]
 Random points from KMM cluster 1: [1121, 2640, 268]

cluster	Image 1	Image 2	Image 3
0			
label	Real sea	Fake forest	Fake mountain
1			
label	Real sea	Fake mountain	Real forest

نتایج سه خوشہ •



Random points from KMM cluster 0: [1581, 2403, 403]

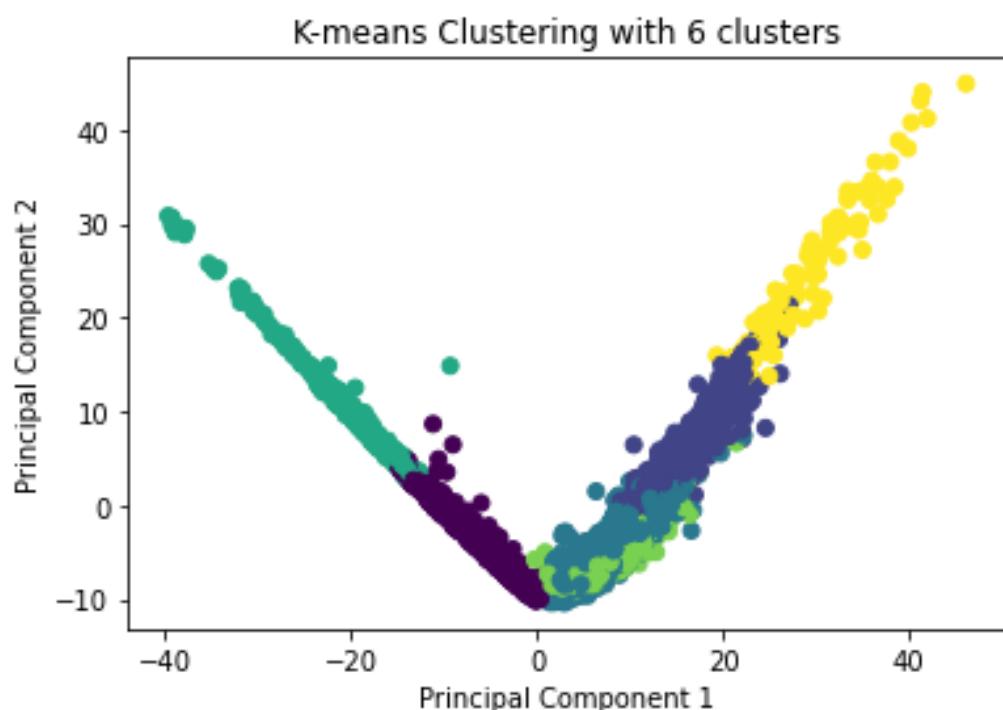
Random points from KMM cluster 1: [224, 1077, 276]

Random points from KMM cluster 2: [2388, 670, 2569]

cluster	Image 1	Image 2	Image 3
0			
Label	Fake forest	Real mountain	Real sea
1			
Label	Fake forest	Real mountain	Fake forest

2			
label	Fake sea	Real sea	Fake mountain

نتائج شش خوشة •

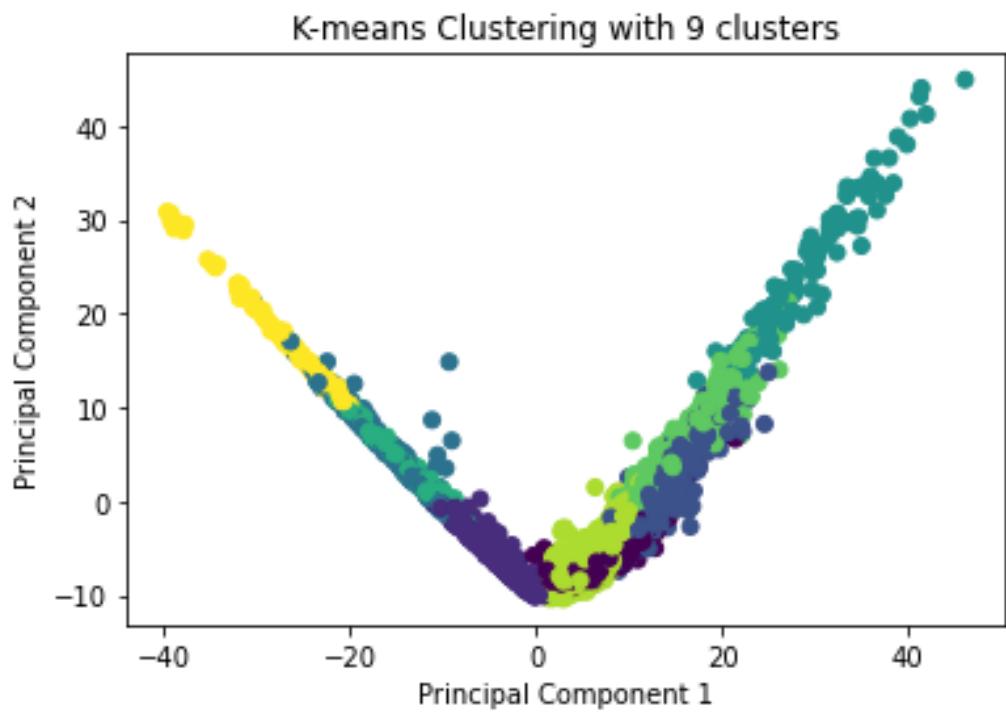


```
Random points from KMM cluster 0: [1887, 2119, 2363]
Random points from KMM cluster 1: [981, 2353, 526]
Random points from KMM cluster 2: [3226, 3318, 22]
Random points from KMM cluster 3: [3029, 397, 1232]
Random points from KMM cluster 4: [1785, 122, 2306]
Random points from KMM cluster 5: [2929, 566, 2118]
```

cluster	Image 1	Image 2	Image 3
---------	---------	---------	---------

0			
Label	Fake mountain	Fake mountain	Real forest
1			
Label	real mountain	real mountain	fake sea
2			
label	Fake sea	Real mountain	fake mountain
3			
label	Real sea	Real sea	Fake forest
4			
Label	Fake forest	Real sea	Fake sea
5			
Label	Real sea	Real forest	Fake forest

نتایج نه خوشہ •

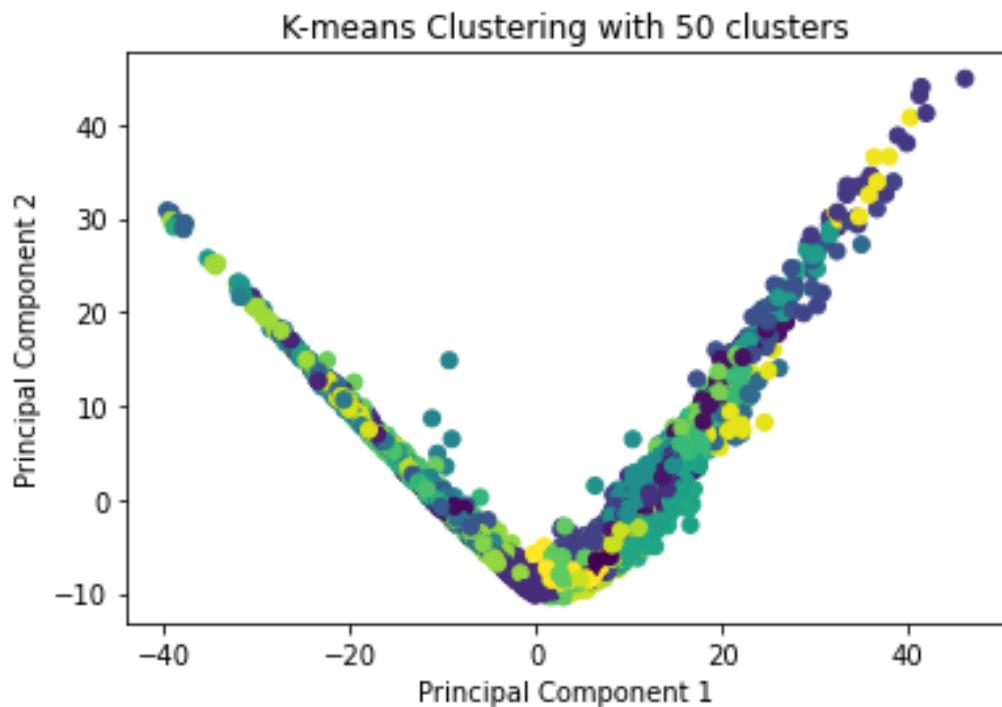


```

Random points from KMM cluster 0: [1319, 2373, 1032]
Random points from KMM cluster 1: [759, 3142, 3102]
Random points from KMM cluster 2: [2890, 2140, 490]
Random points from KMM cluster 3: [2088, 748, 2089]
Random points from KMM cluster 4: [1626, 575, 1236]
Random points from KMM cluster 5: [2941, 879, 1439]
Random points from KMM cluster 6: [2904, 3305, 2096]
Random points from KMM cluster 7: [814, 493, 465]
Random points from KMM cluster 8: [1111, 1560, 2093]

```

cluster	Image 1	Image 2	Image 3
0	Real mountain	Fake mountain	Fake mountain
1	Real forest	Real mountain	Fake forest
2	Real mountain	Fake forest	Real mountain
3	Fake mountain	Real sea	Fake forest
4	Fake sea	Real sea	Fake mountain
5	Real mountain	Real mountain	Real forest
6	Fake sea	Real mountain	Real sea
7	Fake mountain	Fake mountain	Real forest
8	Fake forest	Real sea	Fake forest



```

Random points from KMM cluster 0: [3108, 1851, 560]
Random points from KMM cluster 1: [1902, 3021, 1578]
Random points from KMM cluster 2: [1175, 1972, 2426]
Random points from KMM cluster 3: [761, 2586, 2687]
Random points from KMM cluster 4: [825, 2275, 1987]
Random points from KMM cluster 5: [2748, 1638, 2305]
Random points from KMM cluster 6: [2325, 1584, 881]
Random points from KMM cluster 7: [1942, 305, 3178]
Random points from KMM cluster 8: [1181, 2496, 1341]
Random points from KMM cluster 9: [65, 3294, 166]
Random points from KMM cluster 10: [2367, 1041, 1369]
Random points from KMM cluster 11: [751, 1703, 348]
Random points from KMM cluster 12: [1199, 1031, 1926]
Random points from KMM cluster 13: [1894, 1737, 1439]
Random points from KMM cluster 14: [247, 2999, 3046]
Random points from KMM cluster 15: [104, 1547, 3122]
Random points from KMM cluster 16: [726, 1766, 3060]
Random points from KMM cluster 17: [2045, 1467, 110]
Random points from KMM cluster 18: [3052, 2430, 526]
Random points from KMM cluster 19: [2337, 2954, 3099]
Random points from KMM cluster 20: [259, 22, 13971]

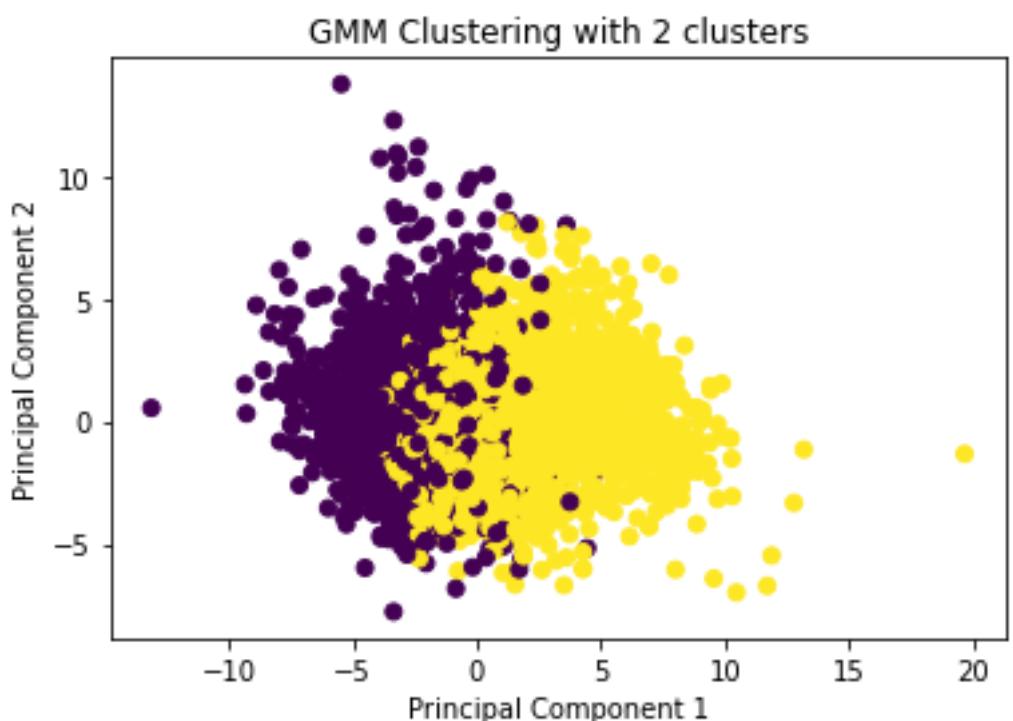
```

3.2- مدل برای ویژگی های استخراج شده

در این مدل ضریب Silhouette از مقدار 0.12 برای دو خوشه تا مقدار 0.08 برای 50 خوشه کاهش می یابد.

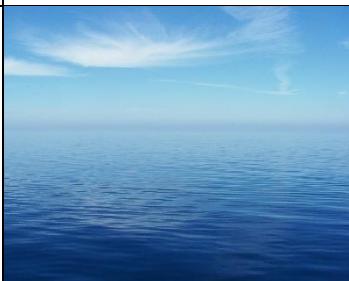
:GMM

نتایج دو خوشه

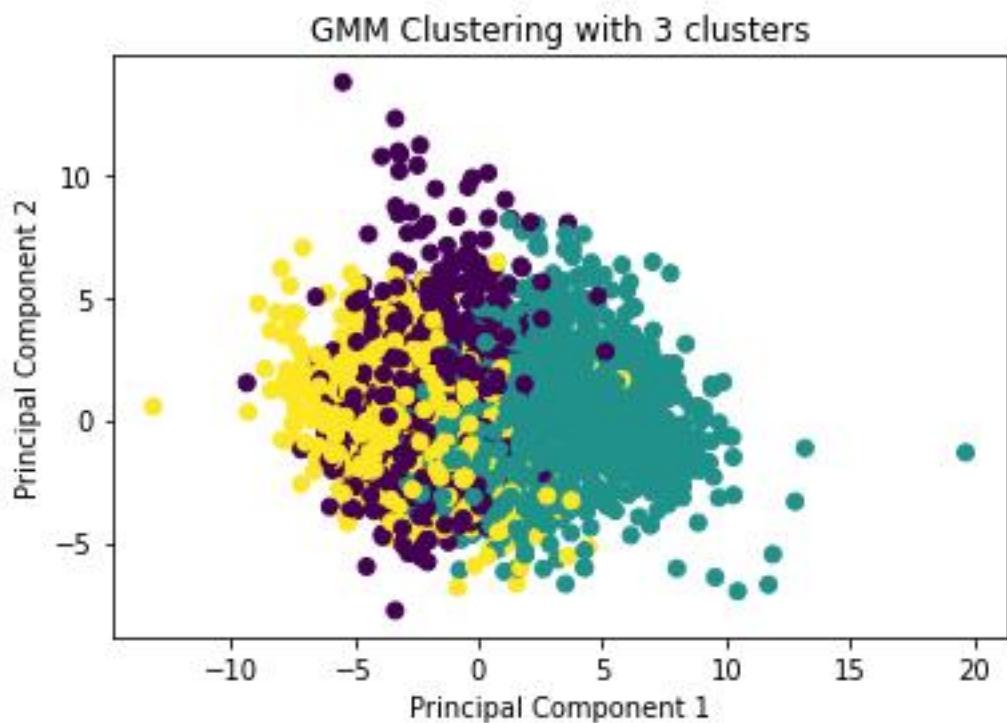


نقاط تصادفی به صورت زیر است

```
Random points from GMM cluster 0: [6, 485, 1461]  
Random points from GMM cluster 1: [2199, 2807, 1974]
```

cluster	Image 1	Image 2	Image 3
0			
label	Real forest	Fake sea	Fake mountain
1			
label	Real sea	Fake sea	Fake sea

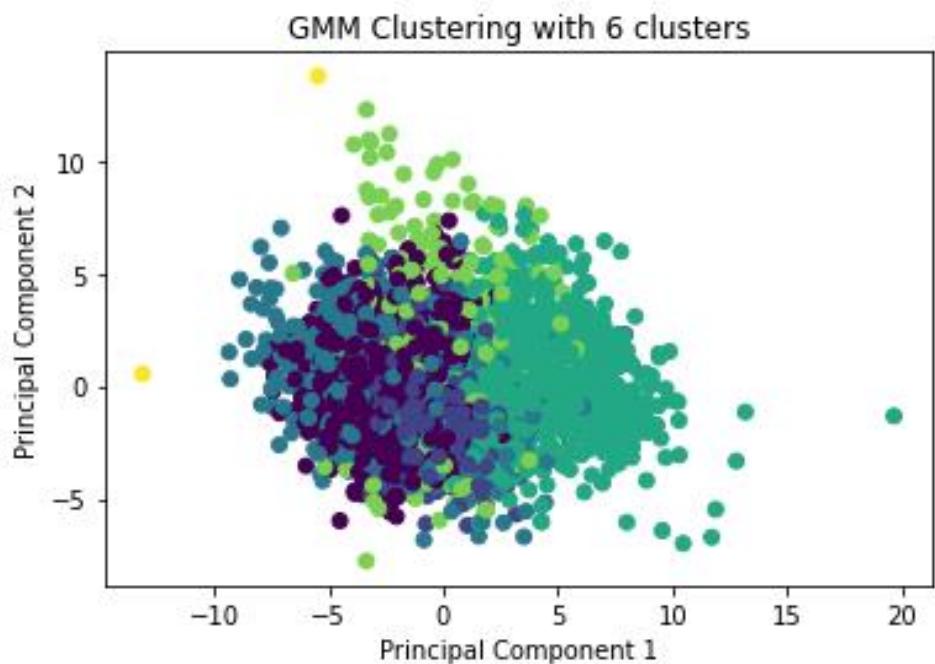
نتایج ۳ خوشہ •



```
Random points from GMM cluster 0: [1273, 2398, 3078]
Random points from GMM cluster 1: [1234, 1887, 37]
Random points from GMM cluster 2: [1731, 662, 400]
```

cluster	Image 1	Image 2	Image 3
0			
Label	Real sea	Real forest	Real maountain
1			
Label	Fake forest	Fake mountain	Real sea
2			
label	Fake forest	Real mountain	Real sea

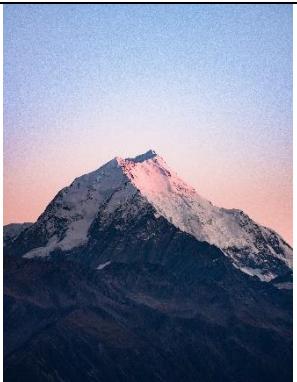
• نتایج ۶ خوش



```

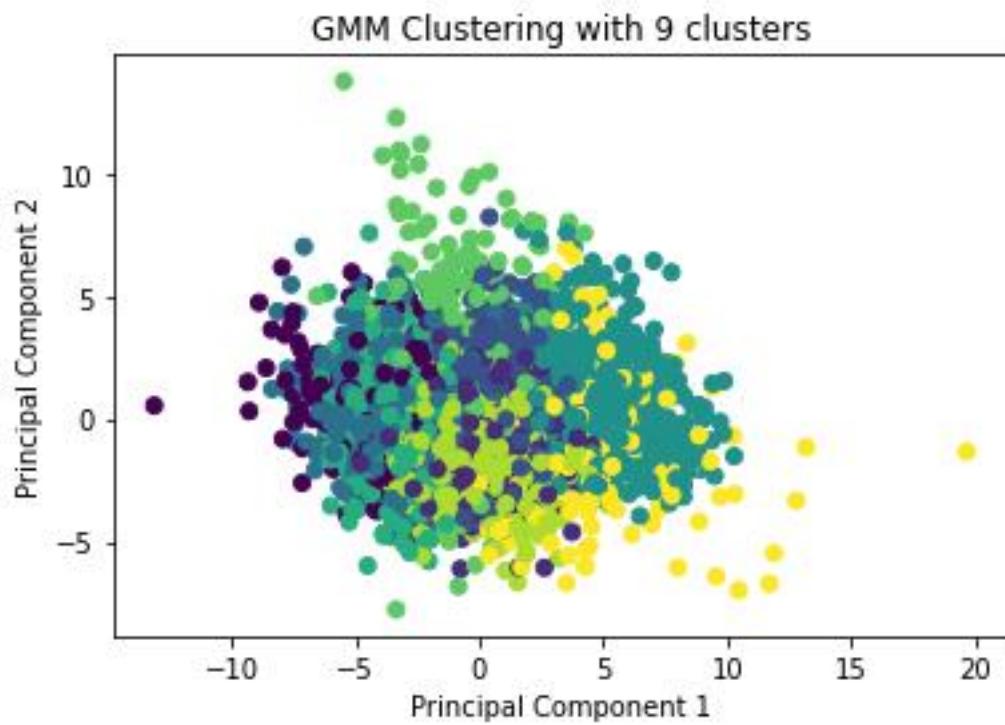
Random points from GMM cluster 0: [2444, 1649, 1498]
Random points from GMM cluster 1: [3141, 762, 628]
Random points from GMM cluster 2: [1023, 1713, 1951]
Random points from GMM cluster 3: [1028, 2126, 2819]
Random points from GMM cluster 4: [1490, 3043, 1316]
Random points from GMM cluster 5: [1260, 394, 3200]

```

cluster	Image 1	Image 2	Image 3
0			
Label	fake sea	Real mountain	Real forest
1			

Label	real forest	real forest	fake sea
2			
label	real sea	Real mountain	fake forest
3			
label	Real sea	Real mountain	Fake mountain
4			
Label	Real mountain	Fake mountain	Fake sea
5			
Label	Real sea	Fake sea	Real sea

• نتایج نه خوش



```

Random points from GMM cluster 0: [3403, 1031, 1874]
Random points from GMM cluster 1: [3158, 123, 1048]
Random points from GMM cluster 2: [86, 1813, 1533]
Random points from GMM cluster 3: [1827, 251, 189]
Random points from GMM cluster 4: [825, 1487, 3156]
Random points from GMM cluster 5: [2033, 2019, 2109]
Random points from GMM cluster 6: [220, 2080, 612]
Random points from GMM cluster 7: [2403, 528, 1142]
Random points from GMM cluster 8: [2488, 457, 1962]

```

cluster	Image 1	Image 2	Image 3
0	Fake forest	Real mountain	Real forest
1	Real mountain	Real sea	Real forest
2	Fake forest	Fake sea	Fake sea
3	Real mountain	Real mountain	Fake sea
4	Fake mountain	Fake mountain	Real sea
5	Fake forest	Fake forest	Fake sea
6	Real mountain	Real mountain	Fake sea
7	Real mountain	Real mountain	Fake sea

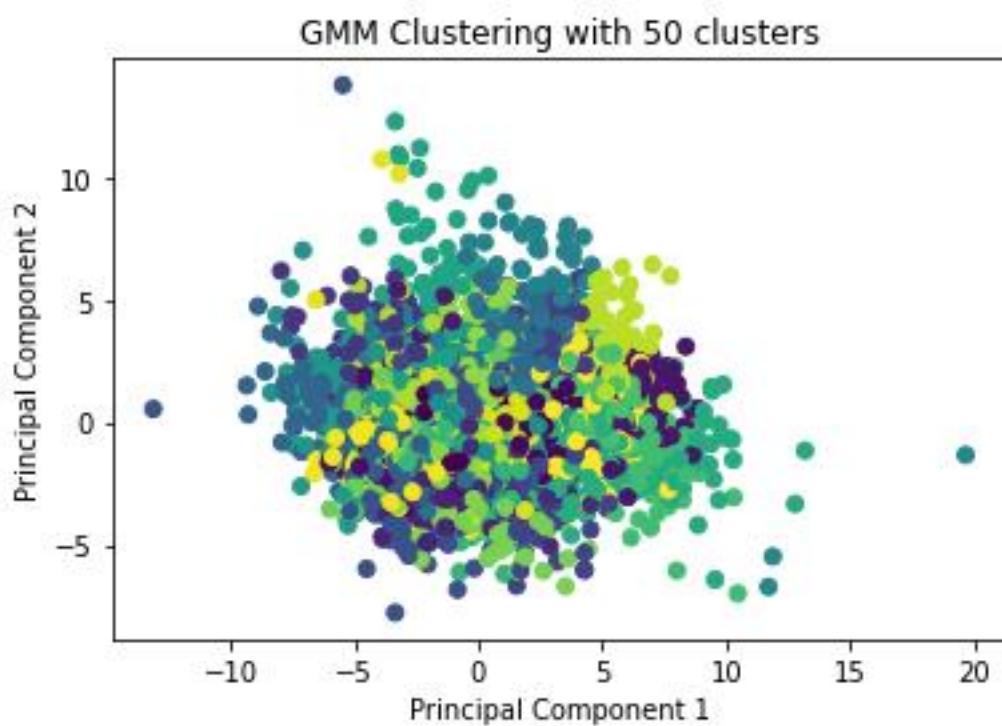
8

Real mountain

Fake mountain

Real forest

نتایج پنجاه خوش



```

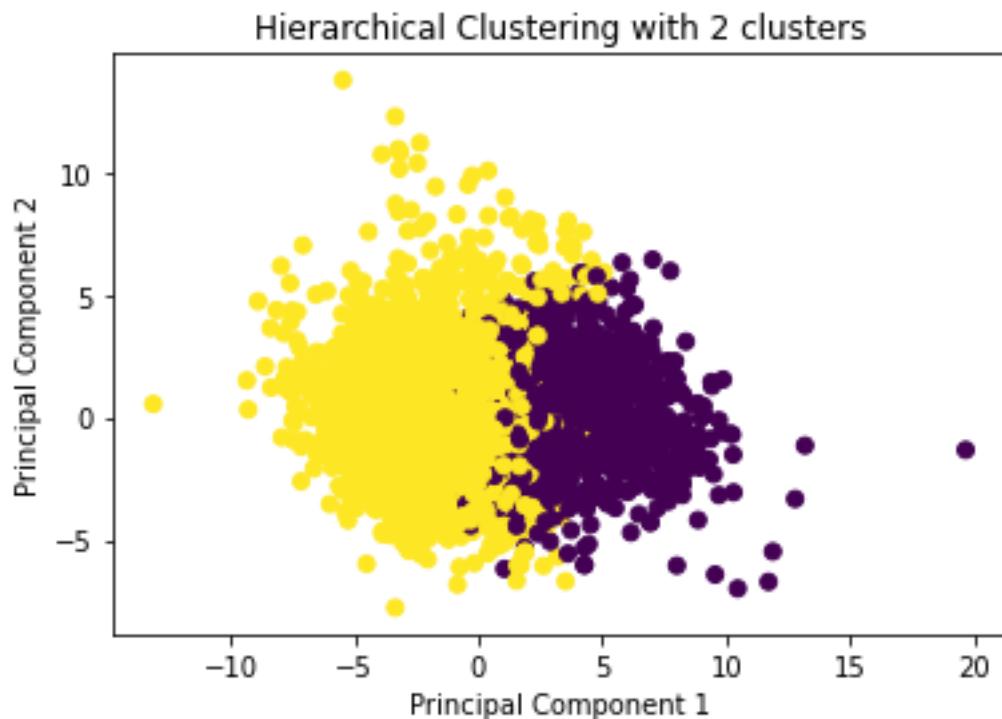
Random points from GMM cluster 0: [1349, 100, 1064]
Random points from GMM cluster 1: [2081, 3249, 2662]
Random points from GMM cluster 2: [920, 42, 326]
Random points from GMM cluster 3: [946, 165, 752]
Random points from GMM cluster 4: [2316, 439, 497]
Random points from GMM cluster 5: [2359, 264, 1709]
Random points from GMM cluster 6: [2783, 2693, 2775]
Random points from GMM cluster 7: [1840, 2599, 691]
Random points from GMM cluster 8: [501, 2767, 2017]
Random points from GMM cluster 9: [3127, 302, 2158]
Random points from GMM cluster 10: [3356, 1729, 3344]
Random points from GMM cluster 11: [3262, 101, 2741]
Random points from GMM cluster 12: [456, 1557, 1494]
Random points from GMM cluster 13: [814, 2068, 2493]
Random points from GMM cluster 14: [3196, 551, 365]
Random points from GMM cluster 15: [1586, 2519, 1550]
Random points from GMM cluster 16: [3230, 779, 27]
Random points from GMM cluster 17: [1549, 1996]
Random points from GMM cluster 18: [446, 770, 3170]
Random points from GMM cluster 19: [396, 885, 1336]

```

Hierarchical

در این مدل ضریب Silhouette از مقدار 0.19 برای دو خوشة تا مقدار 0.06. برای 50 خوشه کاهش می یابد. با توجه به نتایج، این مدل بهتر از GMM برای فیچر های استخراج شده عمل می کند

نتایج دو خوشه •



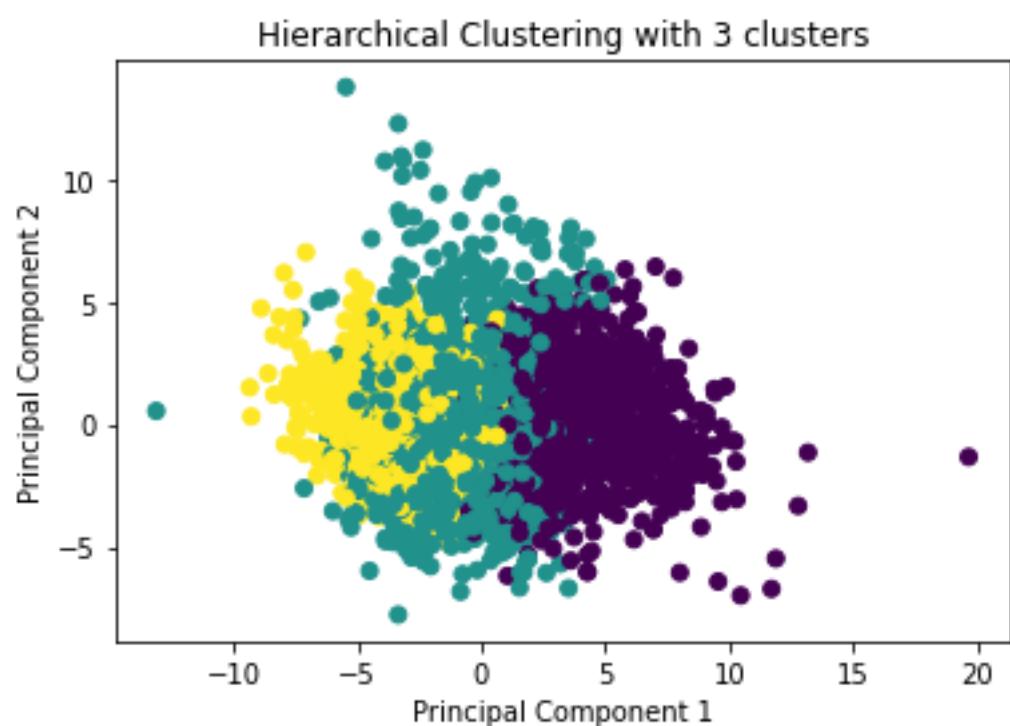
نقاط تصادفی به صورت زیر است.

```
Random points from HAC cluster 0: [980, 3089, 1891]  
Random points from HAC cluster 1: [3377, 27, 3143]
```

cluster	Image 1	Image 2	Image 3
---------	---------	---------	---------

0	Fake sea	Real sea	Fake forest
1			
label	Real forest	Fake sea	Fake mountain

نتایج سه خوشہ •



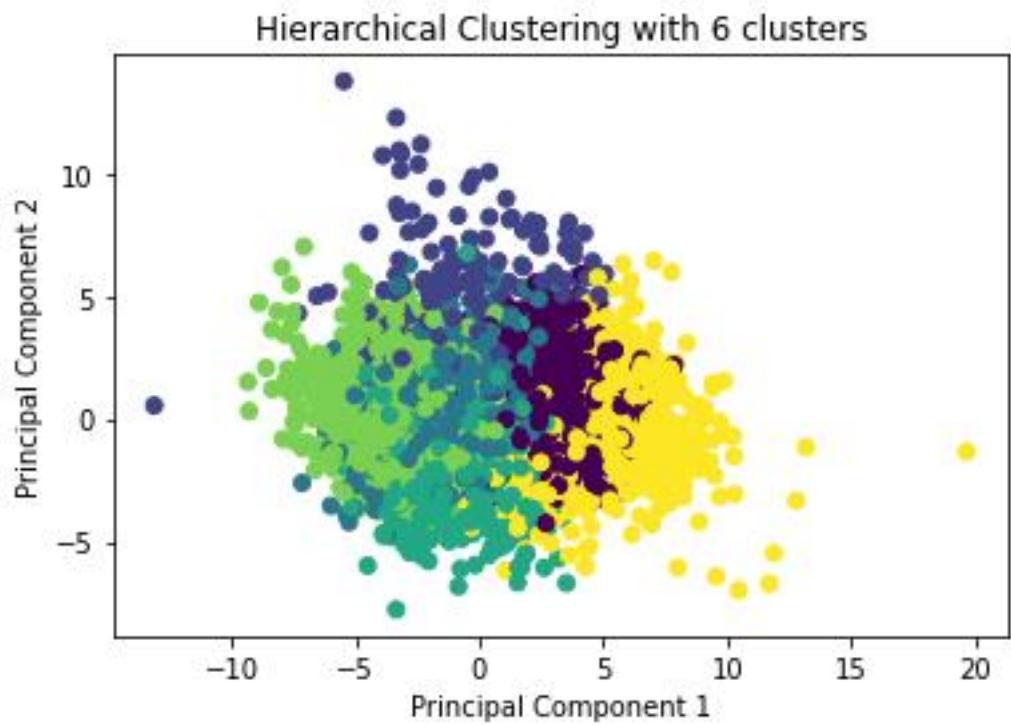
```

Random points from HAC cluster 0: [3126, 1680, 2554]
Random points from HAC cluster 1: [3268, 2080, 2180]
Random points from HAC cluster 2: [3409, 2272, 357]

```

cluster	Image 1	Image 2	Image 3
0			
Label	Fake mountain	Fake sea	Fake maountain
1			
Label	Fake forest	Fake mountain	Real sea
2			
label	Fake forest	Real mountain	Real forest

نتائج شش خوشہ •

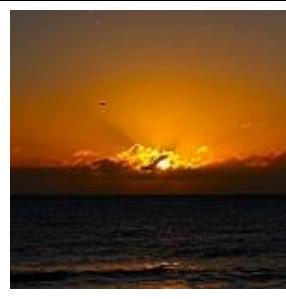


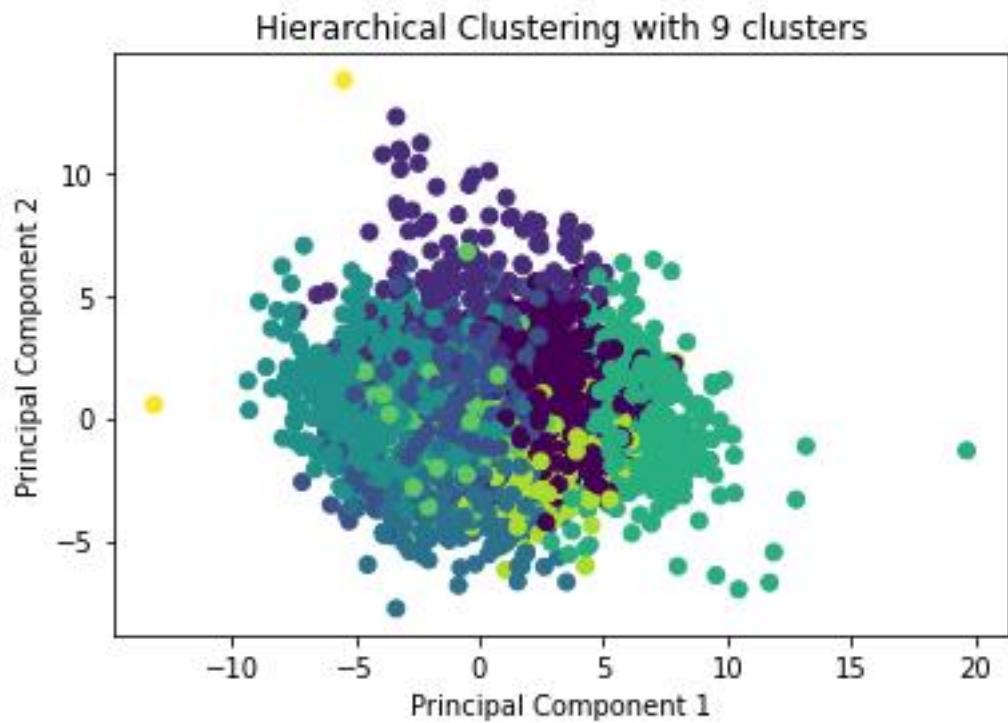
```

Random points from HAC cluster 0: [1722, 2855, 1991]
Random points from HAC cluster 1: [2095, 784, 2037]
Random points from HAC cluster 2: [2886, 1168, 2112]
Random points from HAC cluster 3: [632, 3056, 1625]
Random points from HAC cluster 4: [3144, 1885, 12]
Random points from HAC cluster 5: [1515, 1981, 2408]

```

cluster	Image 1	Image 2	Image 3
0			
Label	fake forest	Real sea	Real sea

1			
Label	real mountain	real sea	fake sea
2			
label	real forest	Real mountain	fake forest
3			
label	Real forest	Fake mountain	Real sea
4			
Label	Fake mountain	Fake mountain	Real mountain
5			
Label	Real forest	Real mountain	Fake sea



```

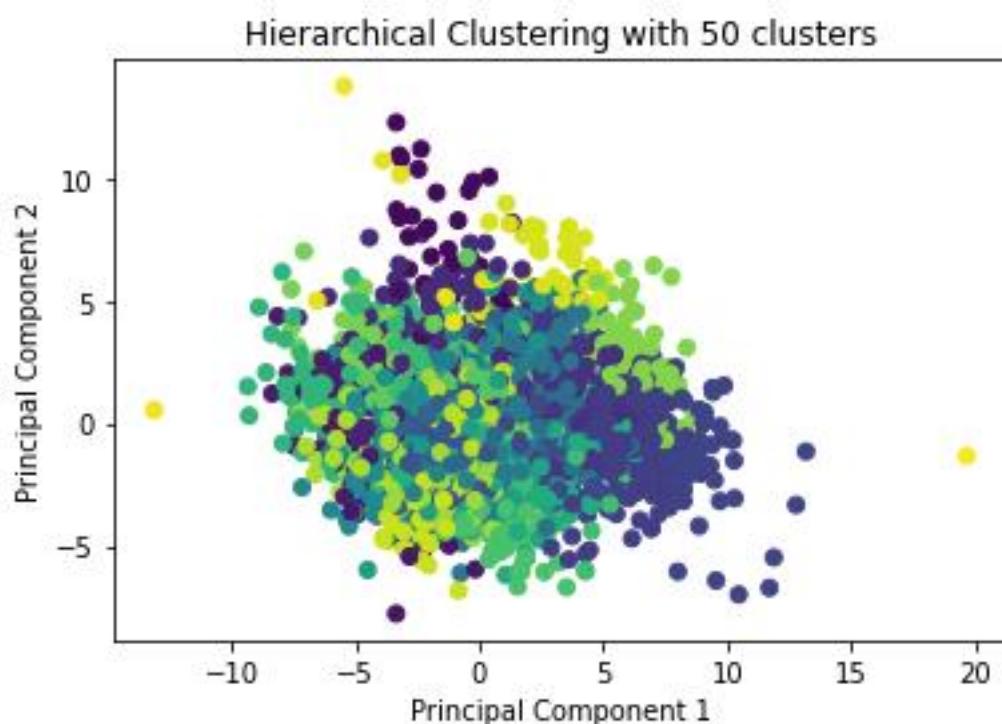
Random points from HAC cluster 0: [1350, 2113, 1745]
Random points from HAC cluster 1: [1248, 1908, 3150]
Random points from HAC cluster 2: [3332, 2949, 3157]
Random points from HAC cluster 3: [2292, 2179, 7]
Random points from HAC cluster 4: [57, 427, 2571]
Random points from HAC cluster 5: [1832, 3298, 2791]
Random points from HAC cluster 6: [477, 1821, 91]
Random points from HAC cluster 7: [2418, 448, 1845]
Random points from HAC cluster 8: [1549, 1996]

```

cluster	Image 1	Image 2	Image 3
0	Real sea	Real forest	Real mountain
1	Real mountain	Fake forest	Real sea
2	Fake forest	Real forest	Fake sea
3	Real sea	Fake sea	Fake forest
4	Fake sea	Fake sea	Fake sea
5	Fake mountain	real forest	Fake mountain
6	Fake sea	Real sea	Fake forest

7	Real mountain	Real sea	Real sea
8	Real forest	Fake forest	

نتایج پنجاه خوشہ •



```

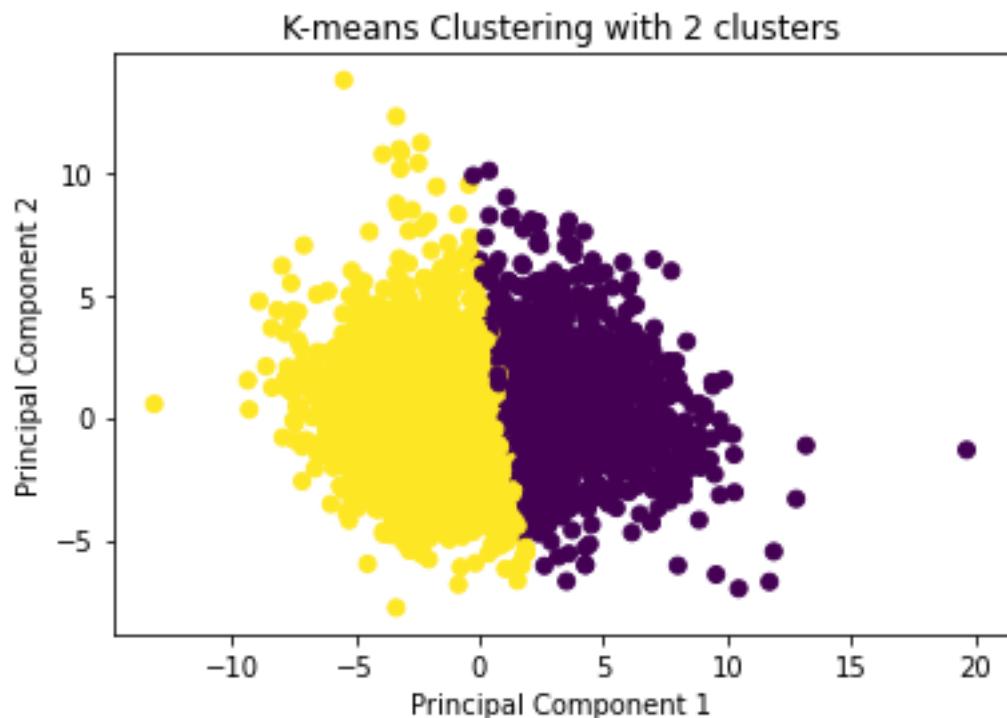
Random points from HAC cluster 0: [511, 693, 121]
Random points from HAC cluster 1: [2908, 2903, 186]
Random points from HAC cluster 2: [2628, 2561, 1584]
Random points from HAC cluster 3: [1805, 492, 35]
Random points from HAC cluster 4: [1121, 2008, 2694]
Random points from HAC cluster 5: [1731, 2420, 2424]
Random points from HAC cluster 6: [461, 104, 3020]
Random points from HAC cluster 7: [1080, 888, 2611]
Random points from HAC cluster 8: [2882, 3038, 2229]
Random points from HAC cluster 9: [2341, 709, 2461]
Random points from HAC cluster 10: [1696, 2948, 2010]
Random points from HAC cluster 11: [1383, 1982, 466]
Random points from HAC cluster 12: [312, 728, 1275]
Random points from HAC cluster 13: [621, 3189, 3252]
Random points from HAC cluster 14: [1816, 2088, 23]
Random points from HAC cluster 15: [822, 25, 1734]
Random points from HAC cluster 16: [2030, 2081, 2219]
Random points from HAC cluster 17: [770, 2727, 597]
Random points from HAC cluster 18: [2798, 1144, 30]
Random points from HAC cluster 19: [1517, 2433, 1278]

```

k-means

در این مدل ضریب Silhouette از مقدار 0.17 برای دو خوشة تا مقدار 0.09 برای 50 خوشه کاهش می یابد.

نتایج دو خوشه •



نقاط تصادفی به صورت زیر است

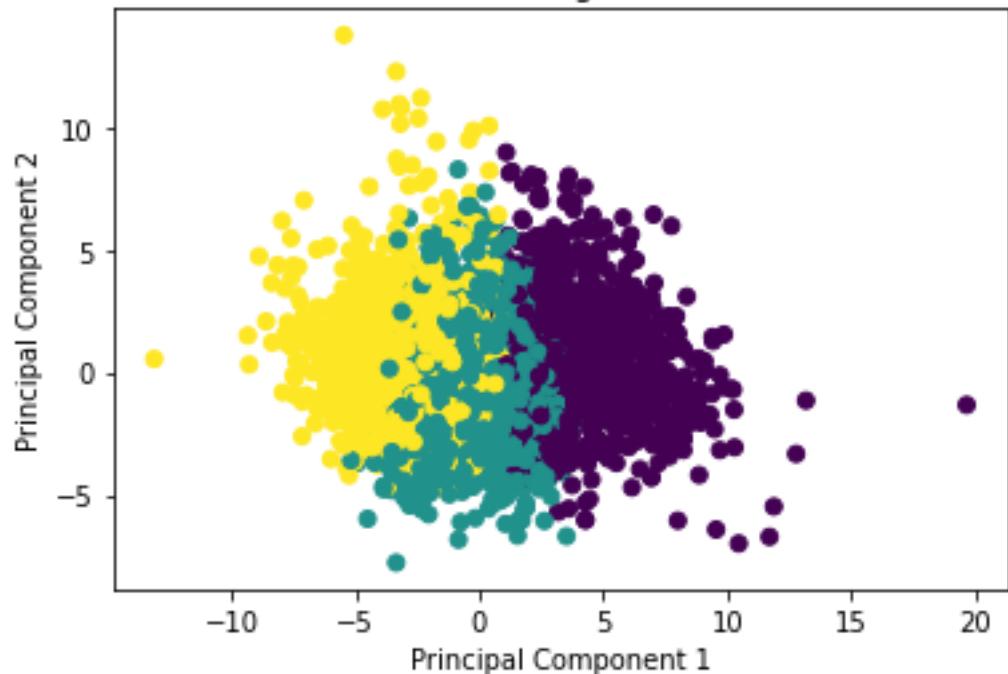
```
Random points from GMM cluster 0: [3222, 1643, 1998]  
Random points from GMM cluster 1: [465, 2193, 360]
```

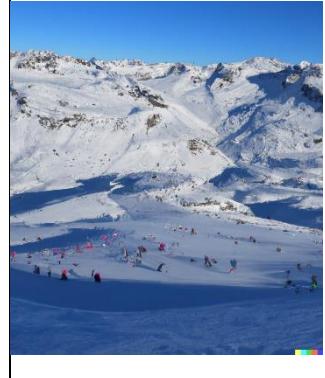
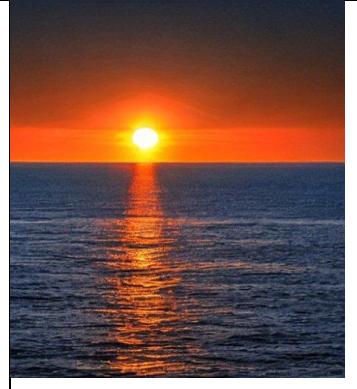
cluster	Image 1	Image 2	Image 3
0			
label	Real forst	Fake mountain	Fake sea

1			
label	Real forest	Fake forest	Fake forest

نتائج سه خوشہ •

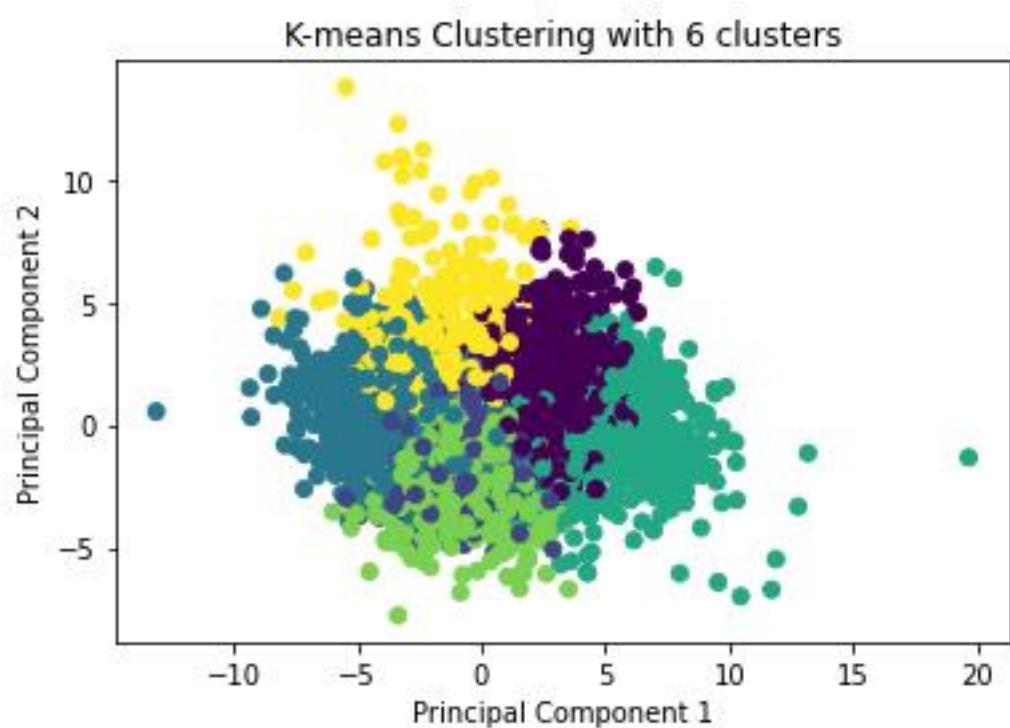
K-means Clustering with 3 clusters



cluster	Image 1	Image 2	Image 3
0			

Label	Real forest	Fake mountain	Fake sea
1			
Label	Fake forest	Real mountain	Fake forest
2			
label	Real sea	Real sea	Fake forest

نتائج شش خوشة •



```

Random points from KMM cluster 0: [287, 3055, 1864]
Random points from KMM cluster 1: [2840, 509, 808]
Random points from KMM cluster 2: [1561, 2769, 962]
Random points from KMM cluster 3: [3131, 2993, 3344]
Random points from KMM cluster 4: [974, 868, 170]
Random points from KMM cluster 5: [3338, 2077, 890]

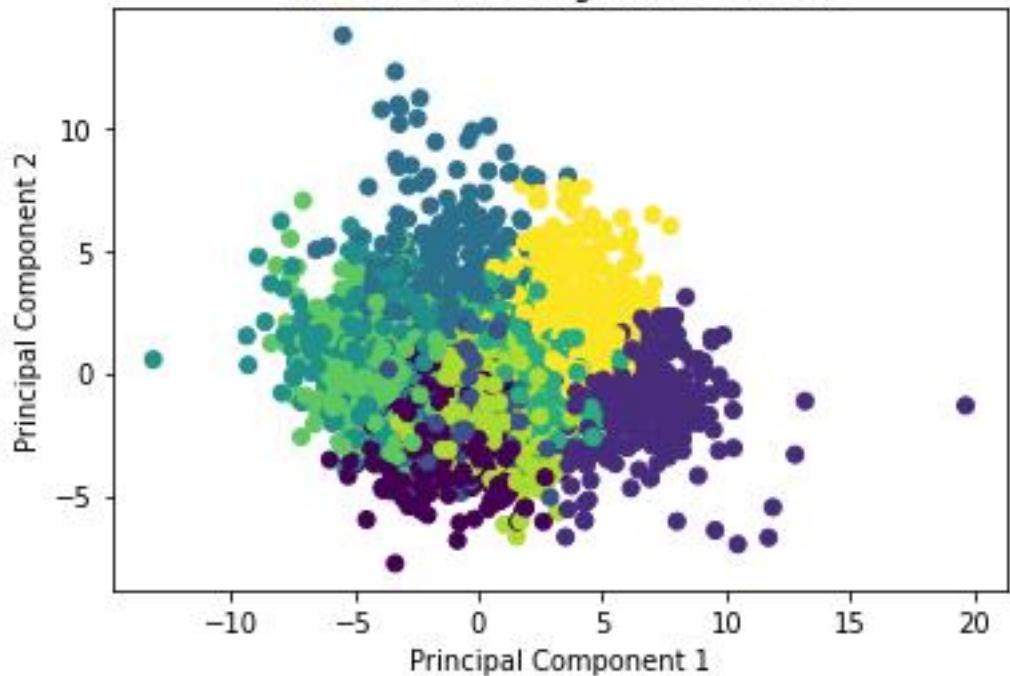
```

cluster	Image 1	Image 2	Image 3
0			
Label	Real forest	Real mountain	fake sea
1			
Label	real forest	real mountain	fake mountain
2			
label	real mountain	fake mountain	fake forest
3			
label	Real forest	Fake forest	Real forest

4			
Label	Fake mountain	Real sea	Real forest
5			
Label	Fake forest	fake mountain	Real forest

نتائج نه خوشہ •

K-means Clustering with 9 clusters

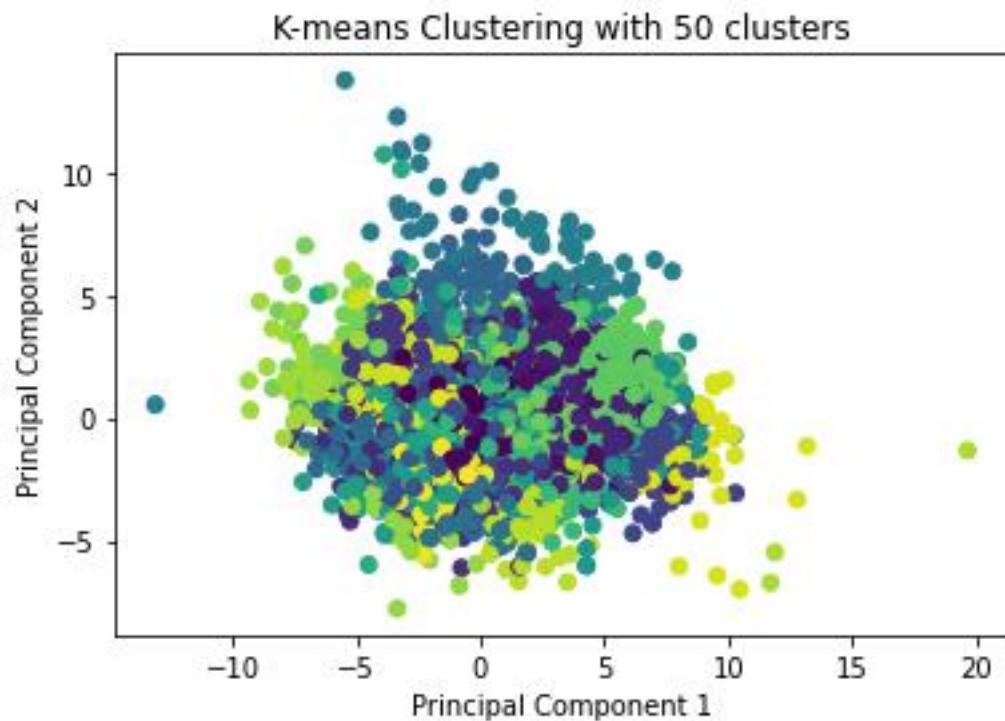


```

Random points from KMM cluster 0: [142, 1323, 3036]
Random points from KMM cluster 1: [1825, 2500, 835]
Random points from KMM cluster 2: [2364, 1850, 883]
Random points from KMM cluster 3: [584, 2276, 1995]
Random points from KMM cluster 4: [245, 2605, 1420]
Random points from KMM cluster 5: [1337, 1770, 184]
Random points from KMM cluster 6: [3248, 3251, 1447]
Random points from KMM cluster 7: [895, 557, 1886]
Random points from KMM cluster 8: [112, 1845, 796]

```

cluster	Image 1	Image 2	Image 3
0	Real sea	Fake mountain	Fake forest
1	Fake sea	Fake forest	Real forest
2	Fake forest	Fake mountain	Fake mountain
3	Fake sea	Fake mountain	Fake forest
4	Real sea	Real forest	Fake mountain
5	Real sea	Real sea	Real sea
6	Real sea	Fake mountain	Fake sea
7	Fake mountain	Fake mountain	Real forest
8	Fake forest	Real sea	Real forest



```

Random points from KMM cluster 0: [2644, 204, 2133]
Random points from KMM cluster 1: [2955, 1095, 406]
Random points from KMM cluster 2: [3315, 1136, 2272]
Random points from KMM cluster 3: [1122, 2392, 1109]
Random points from KMM cluster 4: [1662, 1854, 3109]
Random points from KMM cluster 5: [2556, 2761, 226]
Random points from KMM cluster 6: [420, 95, 1116]
Random points from KMM cluster 7: [1996, 1549]
Random points from KMM cluster 8: [3194, 1747, 2467]
Random points from KMM cluster 9: [1512, 1581, 564]
Random points from KMM cluster 10: [534, 1674, 987]
Random points from KMM cluster 11: [358, 2569, 1734]
Random points from KMM cluster 12: [409, 1444, 1624]
Random points from KMM cluster 13: [527, 19, 407]
Random points from KMM cluster 14: [2869, 1403, 3411]
Random points from KMM cluster 15: [2558, 2613, 870]
Random points from KMM cluster 16: [1981, 1646, 29]
Random points from KMM cluster 17: [2664, 1415, 1123]
Random points from KMM cluster 18: [1255, 1637, 1814]
Random points from KMM cluster 19: [2670, 3101, 2000]
Random points from KMM cluster 20: [2066, 1769, 3136]

```