سلام بر دانشجویان عزیز، چند نکته مهم:

۱. حجم گزارش به هیچ عنوان معیار نمره دهی نیست، در حد نیاز توضیح دهید.

۲. نکته‌ی مهم در گزارش نویسی روشن بودن پاسخ‌ها می‌باشد، اگر فرضی برای حل سوال استفاده می‌کنید حتما آن را ذکر کنید، اگر جواب نهایی عددی است به صورت واضح آن را بیان کنید.

۳. کد های نوشته شده برای هر سوال شبیه سازی را در فایل ipynb متناظر آن سوال بنویسید.

۴. کدهای ارسال شده **بدون گزارش و یا کامنت گذاری دقیق در کد** فاقد نمره می‌باشند.

۵. برای سوالات شبیه سازی، فقط از دیتاست داده شده استفاده کنید.

۶. نمره تمرین از ۱۰۰ نمره می‌باشد

۷. هرگونه شباهت در گزارش و کد مربوط به شبیه سازی، به منزله تقلب می‌باشد و کل تمرین برای طرفین ۱۰۰- خواهد شد.

۸. در صورتی که تشخیص داده شود از چت بات ها به صورت مستقیم برای پاسخ سوال های تئوری و شبیه سازی استفاده شده است، نمره ۱۰۰- در نظر گرفته خواهد شد.

۹. فایل نهایی خود را در یک فایل زیپ شامل، pdf گزارش و فایل کدها آپلود کنید. نام فایل زیپ ارسالی الگوی ML_HW#_StudentNumber داشته باشد.

۱۰. در صورت داشتن سوال، از طریق گروه درس یا ایمیل‌های زیر با تدریسیار مربوطه سوال‌های خود را مطرح کنید.

سوال ۱ و ۲ و ۵ : ehsan.karamii97@gmail.com
سوال ۳ و ۴ و ۶ : amirmfarzane@gmail.com

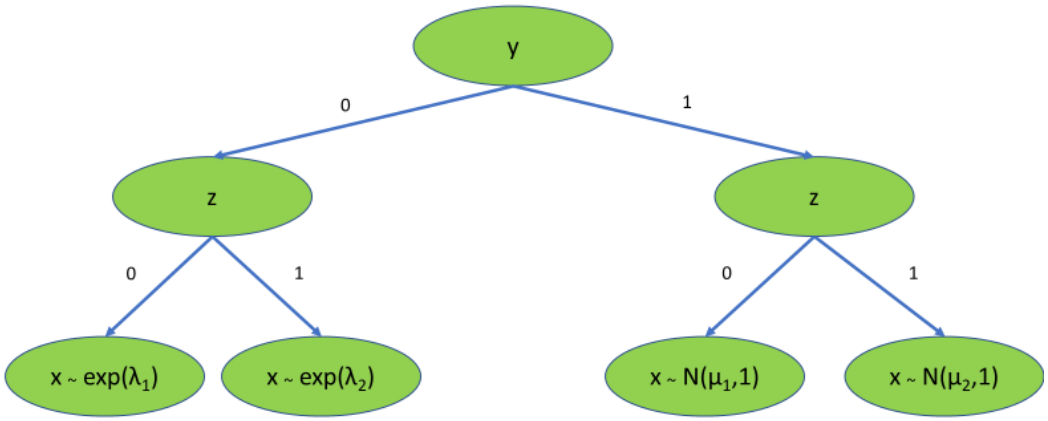
سوال ۱: (۱۰ نمره)

فرض کنید یک مجموعه داده با نمونه هایی مستقل و با توزیع یکسان (i.i.d) از متغیرهای ۱ بُعدی x داریم، و می‌خواهیم از یک شبکه‌ی عصبی چند لایه برای تخمین پارامترهای توزیع مخلوط گوسی بر روی این مجموعه داده استفاده کنیم.

- توضیح دهید خروجی‌های شبکه به چه صورت در نظر گرفته شوند.
- با توجه به مقادیر قابل قبول برای هر یک از پارامترهای توزیع مخلوط گوسی، چه توابع فعالسازی برای خروجی‌های شبکه پیشنهاد می‌کنید.
- تابع هزینه‌ی شبکه به چه صورت خواهد بود.

سوال ۲: (۲۰ نمره)

فرض کنید تولید متغییر تصادفی x، براساس درخت زیر انجام میشود که در آن متغیرهای y و z، باینری و مستقل از یکدیگر هستند.



- اگر احتمال ۱ بودن متغیرهای y و z به ترتیب برابر با α و β باشد؛ آنگاه رابطه توزیع توام سه متغییر x، y وz یا همان P(x,y,z) را بنویسید.
- فرض کنید یک مجموعه داده با نمونه هایی مستقل و با توزیع یکسان (i.i.d) از متغیرهای x تولید شده داریم. تابع لگاریتم complete likelihood را بدست آورید.
- گام E را برای متغییرهای پنهان (y_i و z_i) انجام دهید. توجه کنید که در این گام محاسبه‌ی مقادیر امید ریاضی باید به شرط مشاهده‌ی متغییر x_i محاسبه شوند.
- با استفاده از مقادیر بدست آمده در گام قبلی، مقادیر بهینه‌ی β، μ_2 و λ_2 را بروزرسانی کنید.

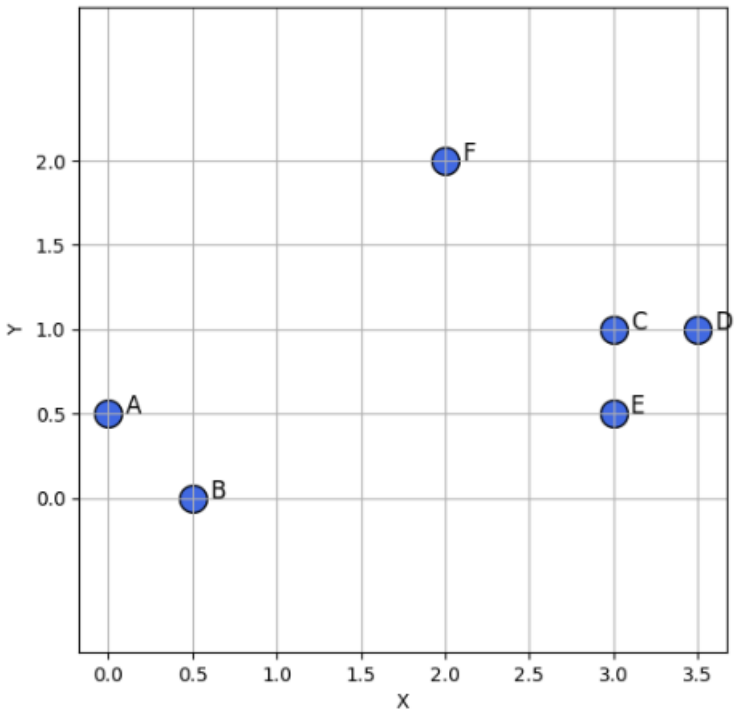
سوال ۳: (۱۰ نمره)

در خوشه‌بندی bottom-up agglomerative clustering ، دو استراتژی رایج برای تعیین فاصله بین خوشه‌ها، یعنی single linkage و complete linkage، به ترتیب حداقل و حداکثر فاصله بین اعضای خوشه‌ها را به عنوان معیار ادغام در نظر می‌گیرند.

الف) پیچیدگی محاسباتی این دو روش را با یکدیگر مقایسه کنید.

ب) کدام روش در برابر outlier مقاوم تر است ؟ پاسخ خود را توجیه کنید.

پ) با استفاده از روش complete linkage خوشه بندی را روی این نقاط انجام داده و سپس نمودار dendrogram را بکشید.

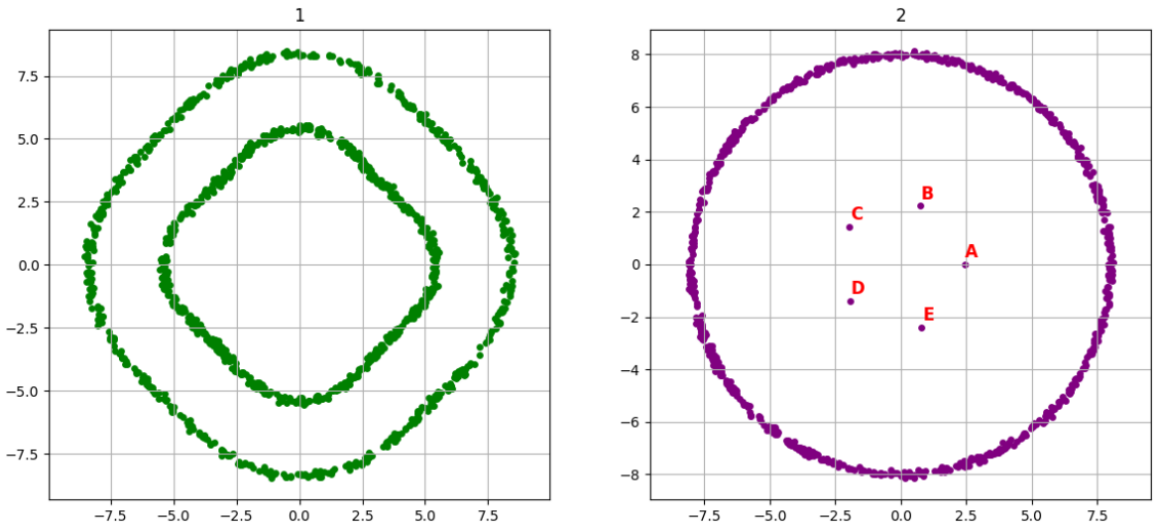


سوال ۴: (۱۰ نمره)

در این سوال به یک روش خوشه بندی مبتنی بر چگالی داده ها یعنی DBSCAN میپردازیم.

الف) به‌طور کلی، در چه شرایطی انتخاب مقدار مناسب برای پارامتر اپسیلون (ε) دشوارتر خواهد بود؟ دو مورد را ذکر کرده و استدلالی منطقی ارائه دهید.

با توجه به نمودارهای ۱ و ۲ می‌خواهیم در هر نمودار تنها دو خوشه داشته باشیم به گونه‌ای که نقاط داخلی و بیرونی به دو خوشه متفاوت تعلق داشته باشند.

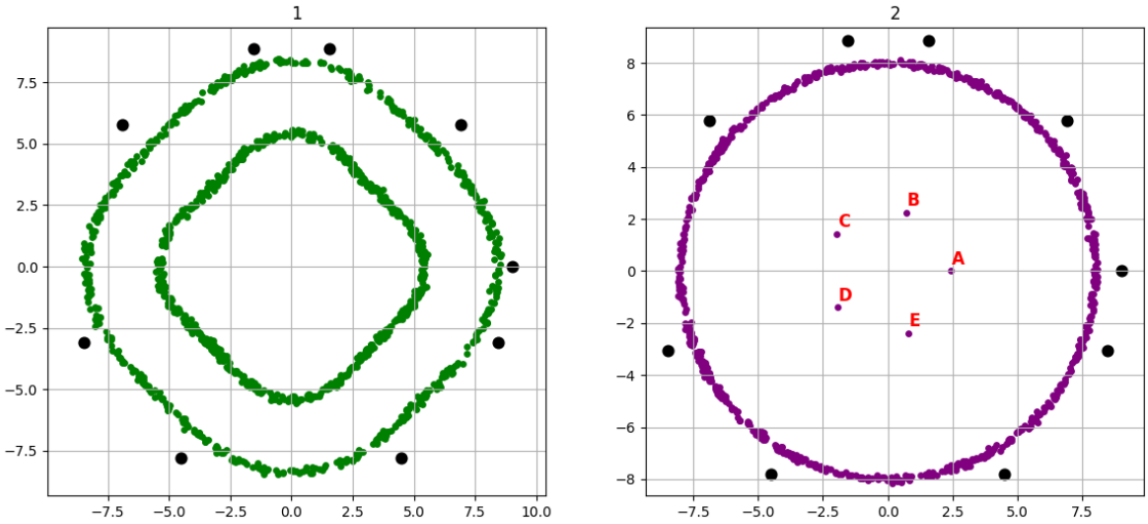


حال به پرسش قسمت ب تا ت پاسخ دهید.

ب) با فرض اینکه مقدار MinPts برابر ۳ انتخاب شده باشد و کمترین اپسیلون را برای تولید این دو خوشه انتخاب کنیم کدام نمودار اپسیلون بزرگ‌تری خواهد داشت؟ دلیل خود را توضیح دهید.

پ) در صورتی که مقدار MinPts برابر ۱ در نظر گرفته شود و کمترین اپسیلون ممکن را انتخاب کنیم ، مقدار آن در مقایسه با حالت بند «ب» چگونه تغییر خواهد کرد؟ برای هر نمودار بررسی کنید و دلیل خود را تحلیل کنید.

ت) در این قسمت تعدادی نقاط سیاه رنگ به اطراف حلقه های بیرونی می افزاییم.



فرض کنید اپسیلون با شرایط گفته شده در قسمت ب محاسبه شده است سپس الگوریتم را ادامه میدهیم تا تکلیف تمام نقاط سیاه نیز مشخص شود.

بررسی کنید در کدام نمودار تعداد بیشتری از نقاط سیاه‌رنگ (outlier) به‌عنوان نویز شناسایی خواهند شد؟ دلیل این تفاوت را با توجه به تراکم نقاط و فواصل آنها از یکدیگر در هر نمودار بیان کنید.

سوال ۵: (شبیه سازی، ۲۰ نمره)

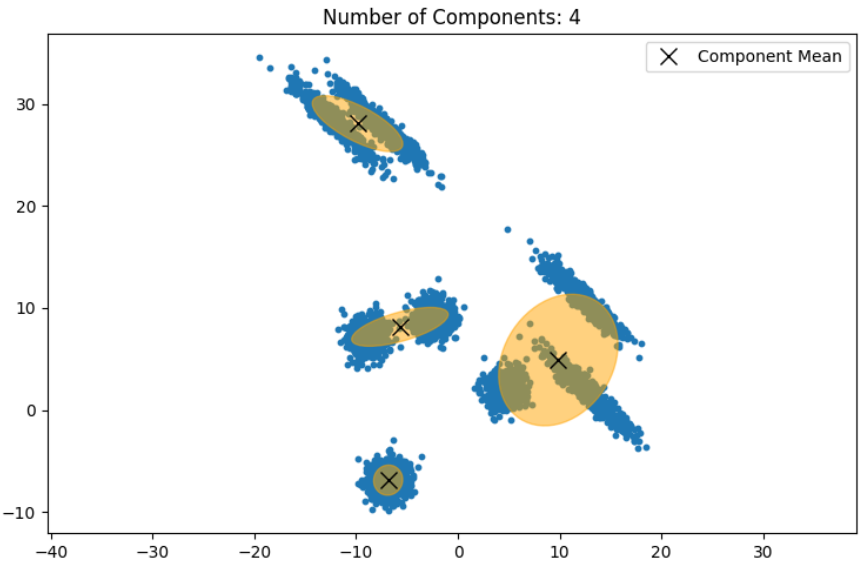
۵-۱- با استفاده از کد زیر، داده مورد نیاز برای انجام این سوال را تولید نمایید.

```
import numpy as np
from sklearn.datasets import make_blobs

n_samples = 2500
random_state = 42
transformation = [[0.6, -0.3], [-0.8, 0.9]]

X, y = make_blobs(n_samples=n_samples, n_features = 2, centers = 4, cluster_std = 1.0,
random_state=random_state)
X_aniso = np.dot(X, transformation) * 2 + 10
x = np.vstack([X, X_aniso])
```

۵-۲- با استفاده از مدل GaussianMixture از کتابخانه sklearn.mixture، توزیع مخلوط گوسی را برای تعداد اجزا (components) در بازه ۱ تا ۱۰ بدست آورید. برای هر مقدار از تعداد اجزا، توزیع‌های بدست آمده را مشابه با نمونه نمایش داده شده در شکل ۵-۱ رسم نمایید. راهنمایی: برای ایجاد بخش‌های زرد رنگ در تصویر زیر، این لینک را مطالعه نمایید.

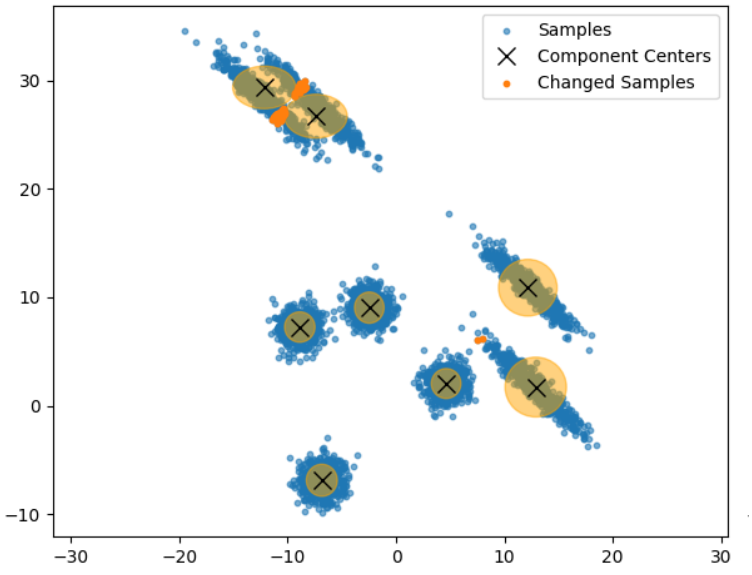


تصویر ۵-۱

۵-۳- مدل‌هایی با تعداد اجزا در بازه‌ی ۱ تا ۲۰ بسازید و مقدار Bayesian Information Criterion (BIC) برای هر مدل را محاسبه نمایید. مقدار BIC را در نموداری بر حسب تعداد اجزا رسم کنید. تعداد اجزای بهینه را مشخص نمایید.

۵-۴- با استفاده از تعداد اجزای بهینه به‌دست آمده در بخش ۲، چهار مدل GaussianMixture با مقادیر مختلف پارامتر covariance_type شامل 'spherical', 'tied', 'diag', 'full' آموزش دهید. برای هر حالت توزیع‌های بدست آمده را مشابه با نمونه نمایش داده شده در شکل ۵-۱ رسم نمایید. تفاوت‌های توزیع‌های بدست آمده را به ازای هر نوع covariance را توضیح دهید.

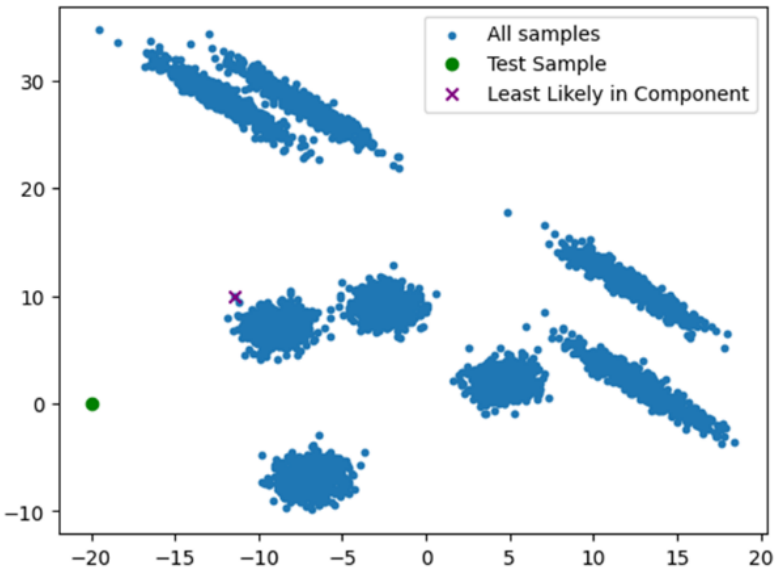
۵-۵- به ازای تعداد اجزای بهینه به‌دست آمده در بخش ۲، با استفاده از دو حالت covariance_type='tied' و covariance_type='full' دو مدل را بر روی داده برازش کنید. نمونه‌هایی را بیابید که برچسب محتمل‌ترین جزء آن‌ها در این دو مدل متفاوت باشد. در صورت وجود چنین نمونه‌هایی، آن‌ها را مشابه با شکل ۵-۲ روی داده‌ها نمایش دهید. راهنمایی: دقت نمایید که برچسب اختصاص داده شده به هر جز در دوحالت، یکسان باشد.



تصویر ۵-۲

۵-۶- به ازای تعداد اجزای بهینه و covariance_type='full' مدلی را بر روی داده برازش کنید. سپس داده‌ای با مختصات (۰, ۲۰-) را در نظر بگیرید و محتمل‌ترین جزء برای این نقطه را تعیین کنید. سپس، تمام داده‌هایی را پیدا کنید که همان برچسب محتمل‌ترین جزء را از مدل گرفته‌اند. برای این داده‌ها، مقدار log-likelihood را با تابع score_samples محاسبه نمایید. نمونه‌ای با کمترین -log likelihood را پیدا کرده و همراه با نقطه (۰, ۲۰-) روی توزیع داده‌ها مشابه با شکل ۵-۳ مشخص

نمایید. مقدار خروجی توابع predict_proba و score_samples را برای هر دو نقطه محاسبه و با هم مقایسه نمایید.



تصویر ۵-۳

سوال ۶: (شبیه سازی، ۳۰ نمره)

الف) در قسمت اول با استفاده از داده‌های خودروها (Q6-cars.csv) ، قصد داریم الگوریتم خوشه‌بندی kmeans را پیاده‌سازی و تحلیل کنیم.

این دیتاست اطلاعات مربوط به خودرو های مختلف است ستون های این دیتاست به این شرح است

- Mpg: نشان‌دهنده مصرف سوخت خودرو بر حسب مایل بر گالن است.
- cylinders :تعداد سیلندرهای موتور را مشخص می‌کند.
- Displacement: حجم موتور را به اینچ مکعب نشان می‌دهد.
- horsepower: قدرت موتور را به اسب بخار بیان می‌کند.
- weight: وزن خودرو را بر حسب پوند نمایش می‌دهد.
- acceleration: مدت زمان شتاب‌گیری خودرو از ۰ تا ۶۰ مایل بر ساعت را به ثانیه نشان می‌دهد.
- Model year: سال ساخت خودرو را نشان می‌دهد (مثلاً ۷۰ به معنی سال ۱۹۷۰ است).
- origin: منطقه ای که خودرو ساخته است مشخص می کند.
- Car name: نام کامل خودرو را به صورت متنی ارائه می‌دهد.

هدف اصلی این است که داده‌ها را به صورت مناسبی پیش‌پردازش کرده، خوشه‌بندی کنید و سپس بر اساس معیار های ارزیابی بهترین تعداد خوشه را تعیین کرده و با کاهش ابعاد کیفیت معیار های ارزیابی برای انتخاب مقدار k را بررسی کنید.

مراحل این قسمت به شرح زیر است:

۱. پیش‌پردازش داده‌ها:

ابتدا باید داده‌ها را بررسی کنید و مقادیر اشتباه را مدیریت کنید و همچنین ستون car name را

میتوانید حذف کنید. همچنین باید باید ویژگی های مجموعه داده را نرمال کنید. (می توانید از StandardScaler) استفاده کنید.

۲. اجرای الگوریتم خوشه‌بندی:

الگوریتم kmeans را برای تعداد خوشه‌های مختلف (از ۲ تا ۷) اجرا کنید.

۳. روش های ارزیابی silhouette coefficient و Davies–Bouldin index را ابتدا توضیح داده و سپس برای k های مختلف این مقادیر را بصورت نموداری نمایش دهید.

۴. حال روش elbow method را برای k های مختلف اجرا کنید.

۵. بر اساس بخش ۳و ۴ بهترین مقدار k را انتخاب کنید. روش خود را توضیح دهید.

۶. با استفاده از روش PCA و t-sne ، نمونه‌ها را برای چند مقدار مختلف k در فضای دوبعدی رسم کنید تا دید بهتری نسبت به نحوه پراکندگی خوشه‌ها و شباهت نمونه‌ها به دست آورید.

۷. آیا خروجی قسمت ۶ میتواند توجیهی برای انتخاب صحیح k توسط شما باشد؟ توضیح دهید.

ب) در این قسمت با استفاده از الگوریتم kmeans تصویر را بر اساس رنگ‌های غالب بخش‌بندی می‌کنیم. این روش کاربردهای متنوعی مانند فشرده‌سازی تصویر، ساده‌سازی برای پردازش‌های بعدی،و همچنین خلق آثار هنری دیجیتال دارد. در پایان با تغییر تعداد خوشه‌ها، تأثیر آن روی کیفیت بخش‌بندی را بررسی خواهید کرد. مراحل این قسمت به شرح زیر است:

۱. آماده‌سازی داده‌ها برای خوشه‌بندی

ابتدا تصویر Q6-img.png را با کتابخانه مناسب لود کرده و سپس به یک آرایه دوبعدی از پیکسل‌ها تبدیل کنید؛ به طوری که هر پیکسل شامل سه مقدار رنگ (RGB) باشد.

۲. اجرای الگوریتم KMeans به ازای مقادیر مختلف k

تعداد خوشه‌ها را در مقادیر مختلف بمثلا (۲و ۴و۸)تعیین کنید و الگوریتم kmeans را روی داده‌های پیکسلی اجرا نمایید تا مراکز خوشه‌ها و برچسب هر پیکسل تعیین شود.

۳. ساخت تصویر بخش‌بندی شده

هر پیکسل را با مقدار عددی (رنگ) مرکز خوشه خودش جایگزین کنید تا تصویر نهایی با رنگ‌های محدود ایجاد شود.

۴. نمایش تصاویر

تصویر اصلی و تصاویر بخش‌بندی شده به ازای هر مقدار k را در کنار هم نمایش دهید تا تفاوت‌ها به خوبی قابل مشاهده باشد و تاثیر تعداد k را بررسی کنید.

یک نمونه از بخش بندی به کمک kmeans به این صورت است :