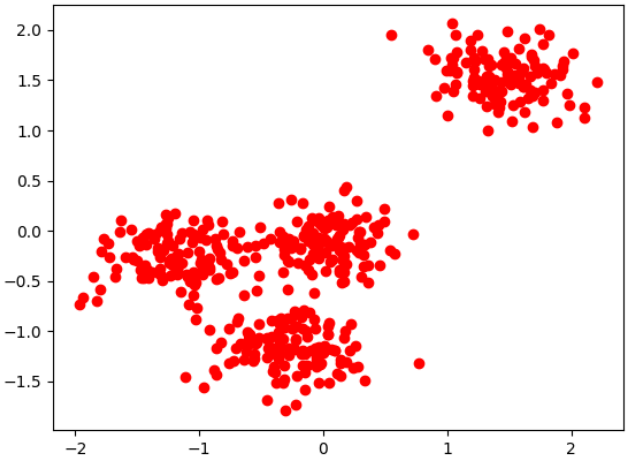
سوال ۱)

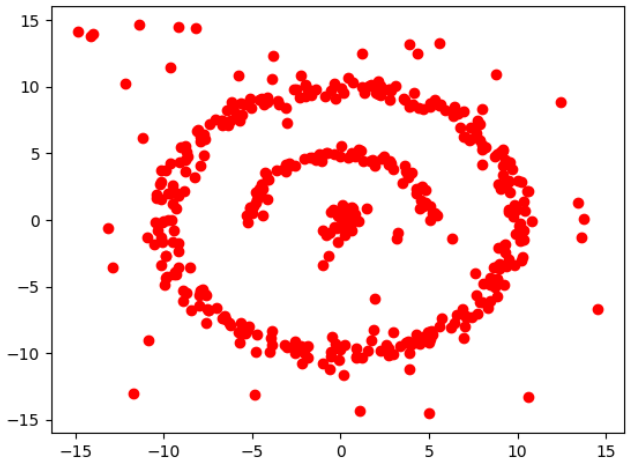
الف)

داده های مصور سازی شده به صورت زیر می باشند.

تصویر زیر متعلق به مجموعه داده Dataset1 می باشد.

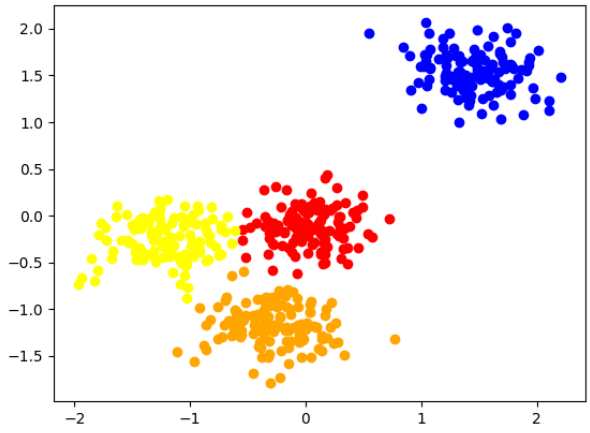


تصویر زیر متعلق به مجموعه داده Dataset2 می باشد.

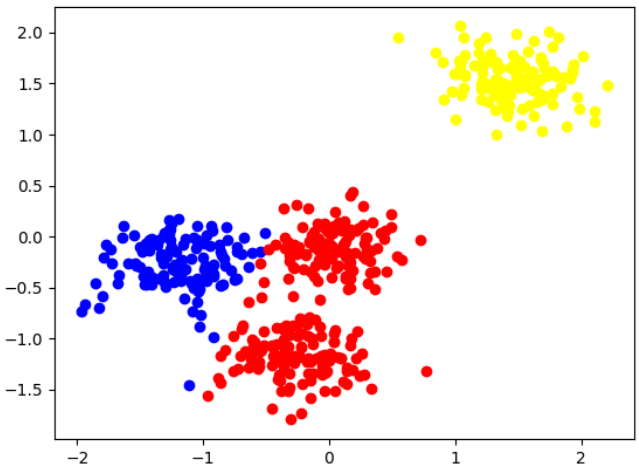


الف) دفعات تکرار الگوریتم برابر با ۱۵ می باشد.

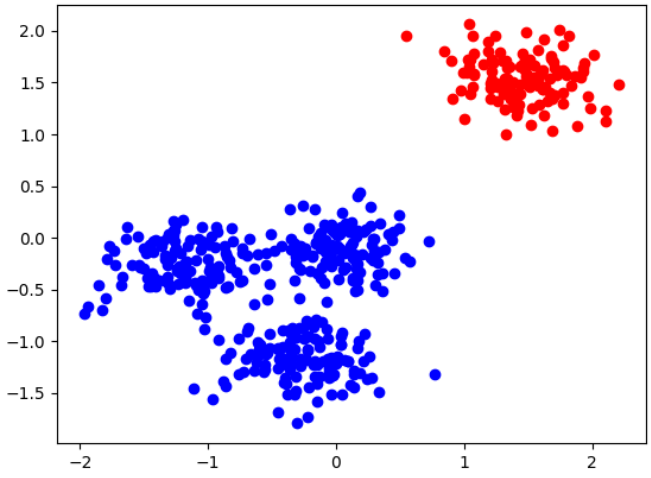
به ازای $k = 4$ تقسیم بندی زیر را داریم.



به ازای $k = 3$ تقسیم بندی زیر را داریم.



به ازای k = 2 تقسیم بندی زیر را داریم.



ب) ترتیب رنگ ها : قرمز، آبی، زرد، نارنجی

خطای خوشه ها به ازای k = 4

0.10684735058049576

0.1330342334241288

0.14440064952159243

0.13259543649909428

خطای خوشه ها به ازای k = 3

0.13259543649909428

0.4237007963735336

0.15221602706905668

خطای خوشه ها به ازای k = 2

0.13259543649909428

0.6254515442772773

ج)

خطای خوشه بندی به ازای k = 4

0.12921941750632782

خطای خوشه بندی به ازای k = 3

0.23617075331389484

خطای خوشه بندی به ازای k = 2

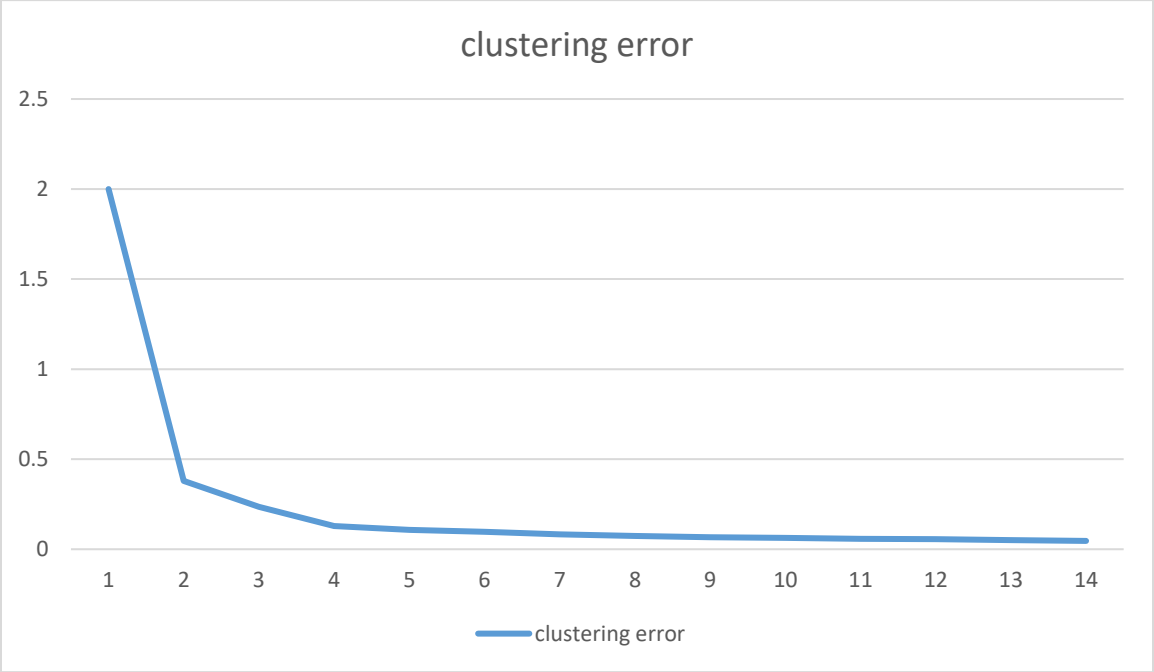
0.37902349038818584

د)

k	Clustering error
1	1.9999999999999991
2	0.37902349038818584
3	0.23617075331389484
4	0.12921941750632782
5	0.10735735275561545
6	0.09639436797129985
7	0.08209671318311854
8	0.0741189891672541
9	0.0668917084286153
10	0.06294479031946285

11	0.056942399436404 97
12	0.055640252208122847
13	0.05035384755097084
14	0.04619485843482697

نمودار منتاظر با جدول بالا به صورت زیر می باشد.



ه)

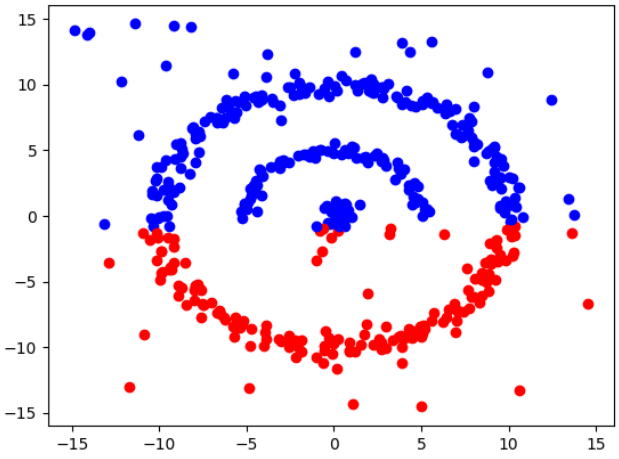
در نمودار دو شکست دیده می شود یکی در $k = 2$ و دیگری در $k = 4$

با توجه به اینکه پس از $k = 2$ هنوز مقدار خطای خوشه بندی کاهش به نسبت زیادی دارد، پس این مقدار برای k مناسب نیست.

بنابراین مقدار k بهینه ۴ می باشد.

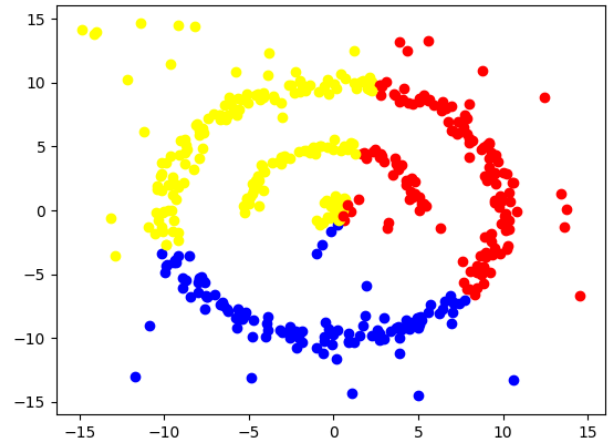
ی)

به ازای k = 2 خوشه بندی زیر را داریم.



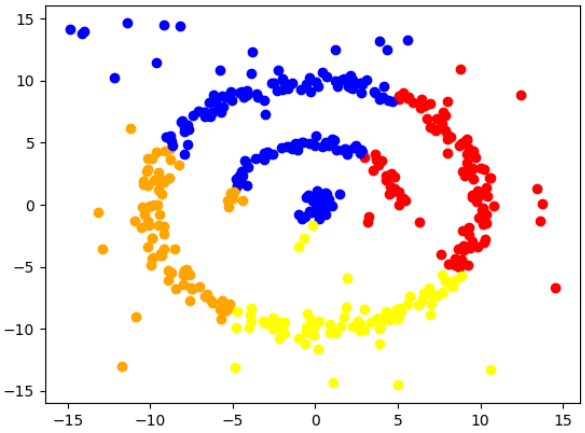
مقدار خطای خوشه بندی در این حالت ۵۶,۲۰۴۹۴۴۳۴۴۵۳۰۴۷ می باشد.

به ازای k = 3 خوشه بندی زیر را داریم.



مقدار خطای خوشه بندی در این حالت ۳۴,۷۸۸۵۸۰۳۲۳۱۰۷۰۰۶ می باشد.

به ازای k = 4 خوشه بندی زیر را داریم.



مقدار خطای خوشه بندی در این حالت ۲۴,۴۴۲۴۱۴۹۱۸۱۶۶۱۴۷ می باشد.

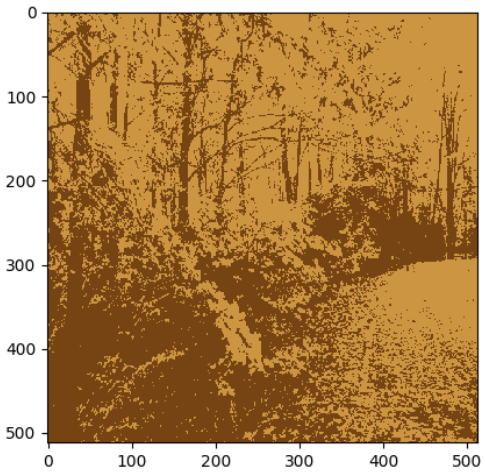
با توجه به حالات بررسی شده و مصور سازی انجام شده و مقدار خطای خوشه بندی در هر حالت مشخصا الگوریتم k-means برای خوشه بندی این داده ها نا مناسب می باشد.

همان طور که در شکل ها دیده می شود در تمامی حالات مرکز خوشه در جایی قرار می گیرد که در حقیقت بین دو بخش مختلفی است که داده های هر کلاس در آن قرار گرفته اند در نتیجه از همه نقاط مربوط به خوشه دور می باشد. این اتفاق باعث می شود مقدار خطای خوشه بندی بسیار زیاد شده و خوشه بندی مان نا مناسب شود.

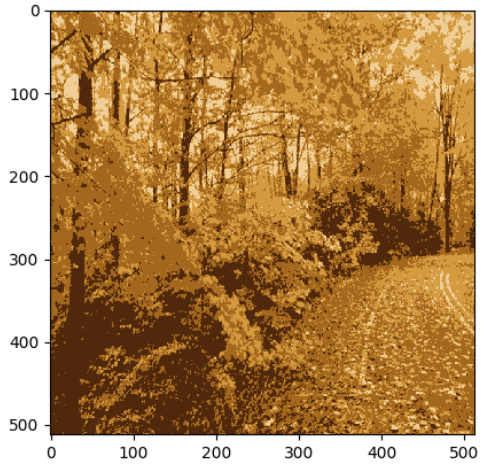
پس می توان گفت الگوریتم k-means برای این مجموعه داده به علت وجود شکل خاص در داده ها نامناسب است و نمی تواند خوشه بندی ای را انجام دهد که صرفا با استفاده از یک مرکز مشخص این نقاط را خوشه بندی کند.

سوال ۲) دفعات تکرار الگوریتم برابر با ۱۰ می باشد.

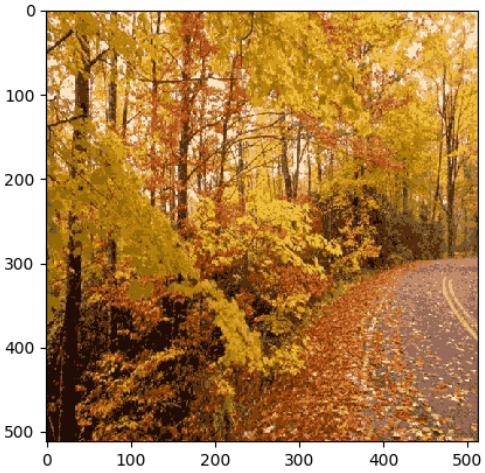
عکس خروجی به ازای k = 2 به صورت زیر می باشد.



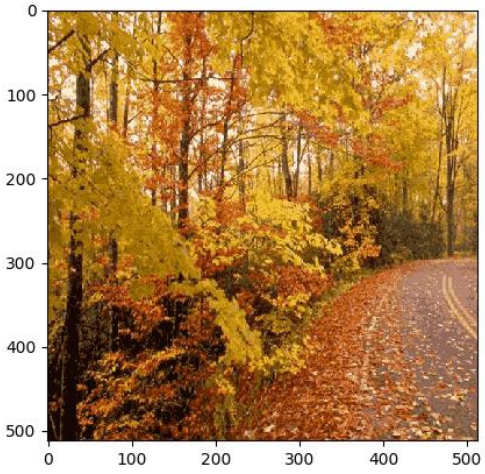
عکس خروجی به ازای k = 4 به صورت زیر می باشد.



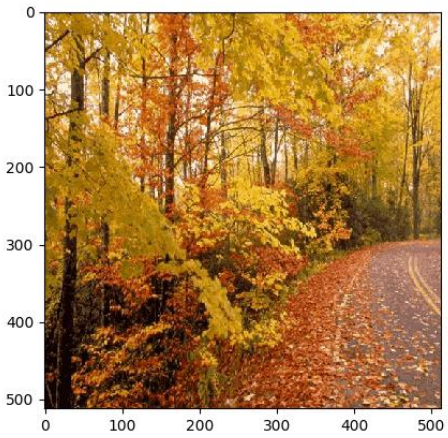
عکس خروجی به ازای k = 16 به صورت زیر می باشد.



عکس خروجی به ازای k = 32 به صورت زیر می باشد.



عکس خروجی به ازای k = 64 به صورت زیر می باشد.



سوال ۳) از داده های covid-sample برای این سوال استفاده شده است.

الف)

نقاط مجموعه داده بر روی نقشه به شکل زیر می باشند.

ب)

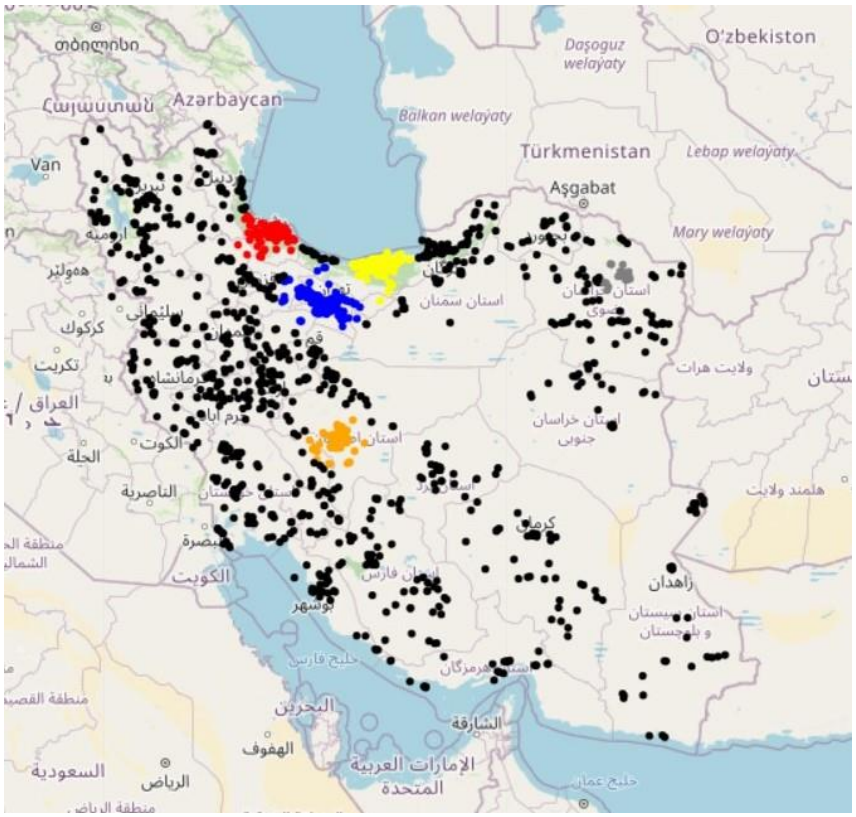
به عنوان یک نمونه با قرار دادن مقدار eps=0.5 و min_sample=20، تعداد خوشه هایمان برابر با ۱۳تا و تعداد داده های نویز نیز برابر با ۵۵۹ تا می شود.

ج)

با تغییر دادن مقادیر eps و min_sample این نتیجه بدست می آید که به ازای eps=0.5 و min_sample=50، ۵ ناحیه متراکم (خوشه) و 1220 داده نویز خواهیم داشت.

د) رنگ های استفاده شده برای خوشه ها به ترتیب قرمز، آبی، زرد، نارنجی، خاکستری می باشد.

شکل بدست آمده به صورت زیر می باشد.



همان طور که در شکل دیده می شود ۵ شهر وجود دارند که تراکم بیماران در آن نقاط زیاد است. این ۵ شهر عبارتند از : تهران، رشت، ساری، اصفهان، مشهد

سوال ۴)

در این سوال از مجموعه

Wikipedia 2014 + Gigaword 5 (6B tokens, 400K vocab, uncased, 300d vectors, 822 MB download)

استفاده شده است.

و سپس در فایل های بدست آمده فایل کم حجم که برای هر کلمه ۵۰ ویژگی در نظر می گیرد، استفاده شده است.

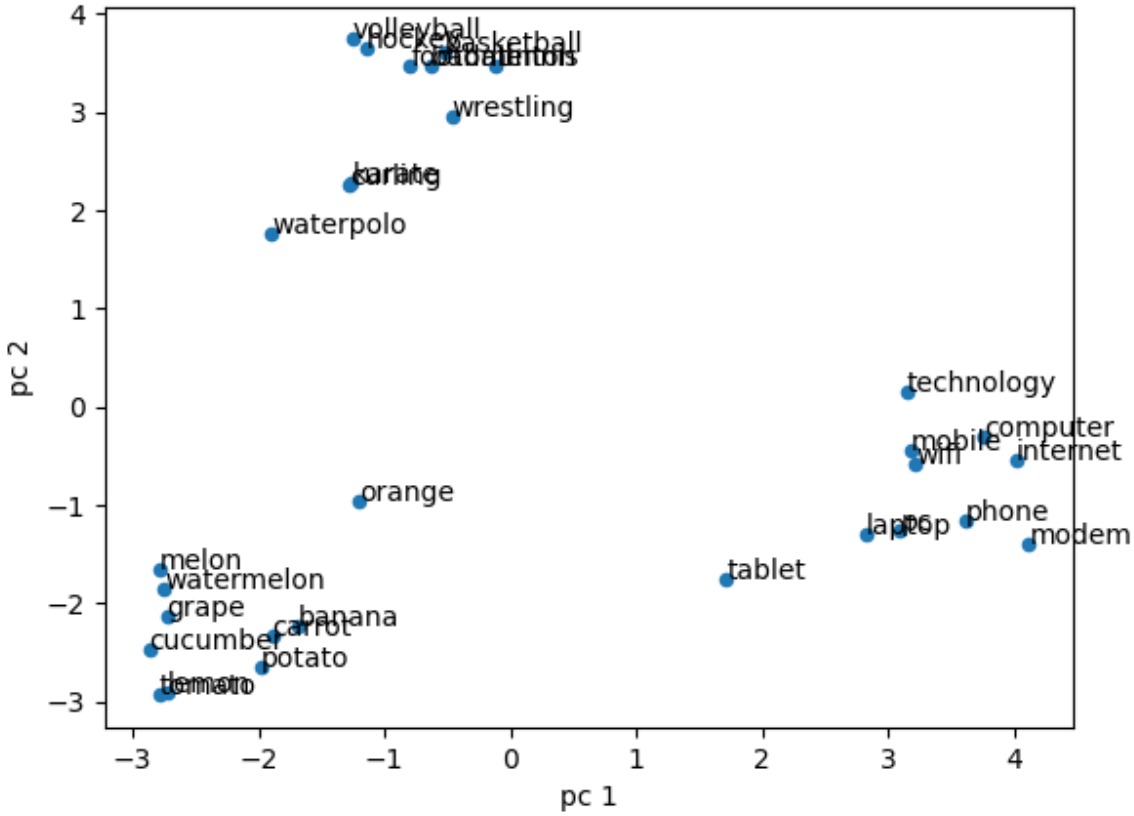
بنابراین الگوریتم PCA داده های ۵۰ بعدی را به داده های ۲ بعدی تبدیل می کند.

این مجموعه دارای ۴۰۰۰۰۱ کلمه می باشد که در مجموعه آن کلمات اعداد و حتی لینک سایت ها نیز قرار دارند.

برای انتخاب زیر مجموعه از این کلمات مجموعه کلماتی را از قبل انتخاب کرده و آن کلمات را به عنوان ورودی به الگوریتم می دهیم.

این کلمات در ۴ گروه قرار دارند و در هر گروه ۱۰ کلمه موجود است. این ۴ گروه عبارت اند از میوه، تکنولوژی، ورزش و کشور.

نتیجه اجرای این الگوریتم برای ۳ گروه اول بیان شده به صورت زیر می باشد.



نتیجه اجرای این الگوریتم برای ۴ گروه بیان شده به صورت زیر می باشد.