

دانشگاه صنعتي امیرکبیر  
(پلی تکنیک تهران)

دانشكده علوم کامپیوتر و ریاضی

پروژه 4 یادگیری ماشین

نگارش

سید حسین محمدی

مدرس

دکتر اکبری

1401

|  |  |
| --- | --- |
| فهرست مطالب | صفحه |

[بخش اول: سؤالات تشریحی 3](#_Toc123158808)

[1-1 انجام k-means بر روی داده 3](#_Toc123158809)

[2-1 PCA و AE 6](#_Toc123158810)

[بخش دوم: تمرین مطالعه (Kernal PCA) 7](#_Toc123158811)

[2-1 مقدمه 7](#_Toc123158812)

[2-2 تعاریف اولیه 7](#_Toc123158813)

[2-3 خاصیت 8](#_Toc123158814)

[بخش سوم: پیادهسازی 9](#_Toc123158815)

[بخش اول: داده Country 9](#_Toc123158816)

[3-1 همبستگی 10](#_Toc123158817)

[3-2 نرمال سازی 10](#_Toc123158818)

[3-3-1 Elbow for k means 11](#_Toc123158819)

[3-3-2 Silhouette explained 11](#_Toc123158820)

[3-3-3 Silhouette for k means 12](#_Toc123158821)

[3-3-4 K means for optimal k 13](#_Toc123158822)

[3-3-5 مقایسه ویژگی ها با تصویر 14](#_Toc123158823)

[3-4-1 Silhouette for GMM 15](#_Toc123158824)

[3-4-2 GMM with optimal k 16](#_Toc123158825)

[3-4-3 مقایسه ویژگی ها با تصویر 17](#_Toc123158826)

[3-4-4 مقایسه GMM با Kmeans 17](#_Toc123158827)

[بخش دوم: Spectral Clustering 18](#_Toc123158828)

[3-5 Custom Spectral Clustering (with kneighbors) 18](#_Toc123158829)

[3-4 ScikitLearn Spectral Clustering (with nearestneighbors) 20](#_Toc123158830)

[3-4 مقایسه با kmeans و GMM 22](#_Toc123158831)

[بخش سوم: PCA 23](#_Toc123158832)

[3-1 موئلفه های مهم 23](#_Toc123158833)

[3-1 مقایسه کلاسترینگ ها 24](#_Toc123158834)

[بخش چهار: داده Shuttle و outlier 27](#_Toc123158835)

[3-2 تشخیص outlier 28](#_Toc123158836)

[بخش پنچم: پیاده سازی دستی 33](#_Toc123158837)

[3-4 KMeans 34](#_Toc123158838)

[3-4 GMM 35](#_Toc123158839)

# بخش اول: سؤالات تشریحی

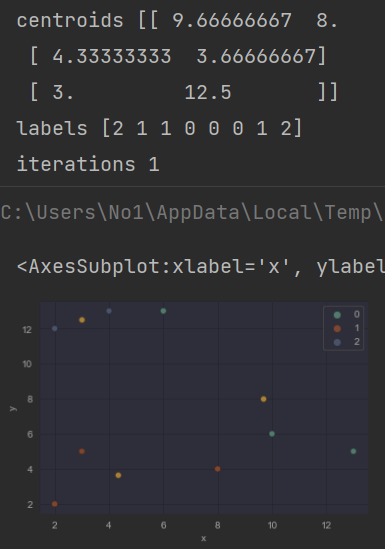
## 1-1 انجام k-means بر روی داده

**بخش الف تا ج:**

حالت اولیه (نقاط طلایی، مرکز ها هستند)

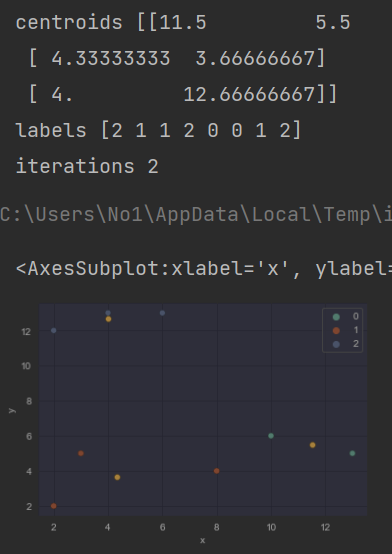


پس از یک قدم



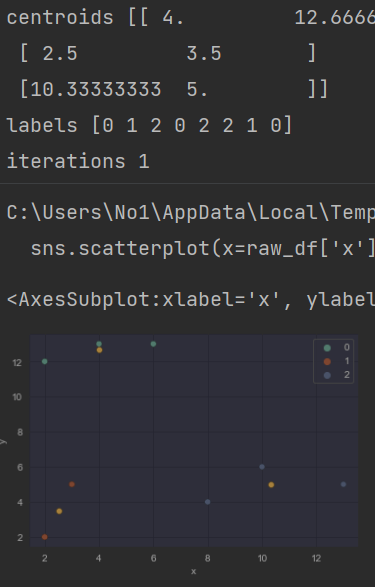
***بخش د:***

*تنها 2 قدم نیاز هست.*



***بخش ه:***

*خیر. نحوه انتخاب نقاط اولیه مرکز تاثیر زیادی بر خوشه بندی میگذارد و با انتخاب رندوم آن ها، میتوان به خوشه بندی بهتری رسید.  
با در نظر گیری آمار هایی مانند silhouette، میتوان فهمید که انتخاب بهتر برای 3 مرکز وجود دارد که در پایین قرار دادیم.*

**

## 2-1 PCA و AE

**بخش الف:**

*PCA به سادگی قابل فهم و انجام هست. اما اساسا فقط یک چرخش در فضا انجام میدهد و نمیتوان از آن انتظار داشت که داده غیر خطی را به خوبی نمایش دهد. باید گفت که kernel pca با کمک از کرنال هایی مانند RBF میتوانند کمی به این مسئله کمک کنند.*

*از طرف دیگر، AE پیچیده تر هست و نمیتوان به راحتی فهمید که به چه شکل کار میکند، اما با لایه های غیر خطی میتواند داده های غیر خطی را فشرده کند. باید اشاره شود که PCA، داده را به صورتی خروجی میدهد که در آن همبستگی وجود ندارد، اما چنین چیزی برای AE لزوما صادق نیست.*

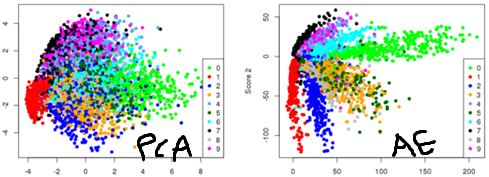
**بخش ب:**

*PCA، زیرا که چرخش در فضا به گونه ای انجام میدهد که همبستگی داده به صفر برسد. اما AE اینچنین نیست.*

**بخش ج:**

*بله. PCA بروی داده یک عملیات خطی (ترانسفورم عمود) انجام میدهد و یک AE با لایه خطی میتواند چنین ترانسفورمی نیز انجام دهد، اما تمرین دادن AE یافتن چنین کانفیگی ممکن هست دشوار باشد.*

**بخش د:**

**

*عکس راستی مربوط به AE هست. زیرا که PCA نمیتواند داده غیر خطی را به خوبی نمایش دهد. همچنین در عکس راستی، تمامی داده نقاط کلاس ها متفاوت با دور شدن از مبدا مختصات، از هم جدا میشوند که اشاره به وجود لایه های RELU دارد. (نزدیکی به مبدا به معنای فعال نشدن نورون هست که مشابه به ناتوانی در تشخیص کلاس هست. به همین دلیل در مبدا، کلاس ها به هم نزدیک هستند)*

# بخش دوم: تمرین مطالعه (Kernal PCA)

## مقدمه

بسیاری از داده ها دارای روابط غیرخطی در بین ویژگی ها هستند. این حقیقت به هنگام استفاده از روش های ترانسفورم (و کاهش بعد) مثل PCA دردسر ساز هستند زیرا که PCA فقط میتواند بر روی داده یک ترانسفورم خطی انجام دهد و به هنگام کاهش بعد، عملکرد خوبی ندارد.

کرنال ها، توابع ای هستند که با ترکیب یک یا چند ویژگی به صورت غیر خطی، تلاش میکنند که داده را "قابل جدا سازی" بکنند.

Kernal PCA، ابتدا داده اصلی ما را به بعد بالا تری میبرد که در آن بعد ممکن هست داده راحت تر قابل تفکیک باشد، سپس با PCA، این داده را به بعد کمتر میبرد.

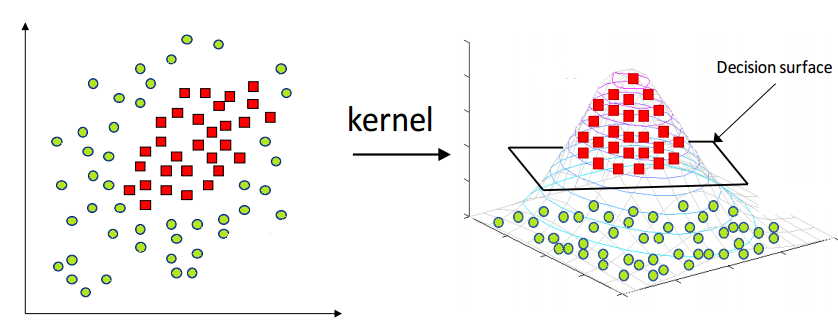
## تعاریف اولیه

ابتدا باید بدانیم که PCA و Kernal چه هستند. در اینجا وارد جزئیات نمیشویم ولی به صورت کلی، با فرض اینکه داده در مرکز باشد (میانگین = 0) PCA یک روش برای یافت یک ترانسفورم (یک ماتریس) برای ماتریس کوواریانس داده هست به شکلی که این کوواریانس بین ویژگی ها به صفر برسد. در عمل این کار باعث یک نوع چرخش داده به دور محور میشود به صورتی که برای ویژگی های جدید (محور های جدید) واریانس هر ویژگی به ماکزیمم شود. (اولین ویژگی بیشترین واریانس و ...).  
معمولا پس از این تغییر محور ها که به صورت قدم به قدم بدست می آیند، محور های پایانی که کمترین واریانس را دارند، حذف میکنند که یک نوع کاهش بعد هست. این ترانسفورم خطی هست (یک چرخش به روز محور) و نمیتوان از آن انتظار داشت که عملکرد خوبی بر روی داده غیر خطی (رابطه غیر خطی بین ویژگی ها) داشته باشد.

کرنال ها توابعی هستند که میتوان آن ها را به صورت زیر نمایش داد.



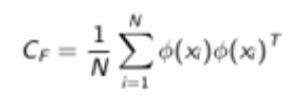
این توابع متشکل از یک نوع ضرب داخلی بین خروجی تابع ای مثل فی هست. این ساختار در بسیاری از مدل ها برای تبدیل داده به یک شکل قابل تفکیک استفاده میشود. برای مثال در SVM، با استفاده از کرنال های مختلف، ویژگی های جدیدی تولید شده که از ترکیب (با کرنال) یک (یا چند) ویژگی دیگر هست. در پایین شکل داده ای را میبینید که نمیتوان با یک هایپرپلین جدا کرد، اما با افزودن یک ویژگی جدید که با کمک از تابع x^2 ساخته شده، میتوان با یک هایپرپلین به راحتی آن را جدا کرد:



## خاصیت

در کل، ابتدا عملیات فی (که کرنال را تعریف میکند)، بر روی داده انجام میدهیم. و داده به شکل جدید تولید میشود، سپس پس از اینکه اطمینان یافتیم که داده در مبدا است (میانگین 0) با انجام PCA کار را تمام میکنیم. برای مثال، داده بالا را نمیتواند در 1 بعد به گونه ای خلاصه کرد که خواص مهم را نگه داشت و چون PCA یک روش unsupervised هست، این کار بسیار دشوار هست. اما پس از عمال کرنال، داده جدید را میتوان در یک بردار به صورت مفید قرار داد.

باید توجه داشت که برای بسیاری از کرنال ها، یافتن تابع فی ممکن نیست و فقط دسترسی به کرنال (ضرب داخلی فی) موجود هست. این امر مشکلی نیست زیرا که PCA در اصل بر روی ماتریس کوواریانس کار میکند که ساخته شده از ضرب مختصات ها در هم هست. به زبان دیگر، میتوان از انجام عملیات فی صرف نظر کرد و فقط تعریف کوواریانس را با مقدار زیر جایگزین کرد تا بتوان kernel pca انجام داد (توجه کنید که ضرب داخلی فی ها، همان کرنال هست که موجود هست):

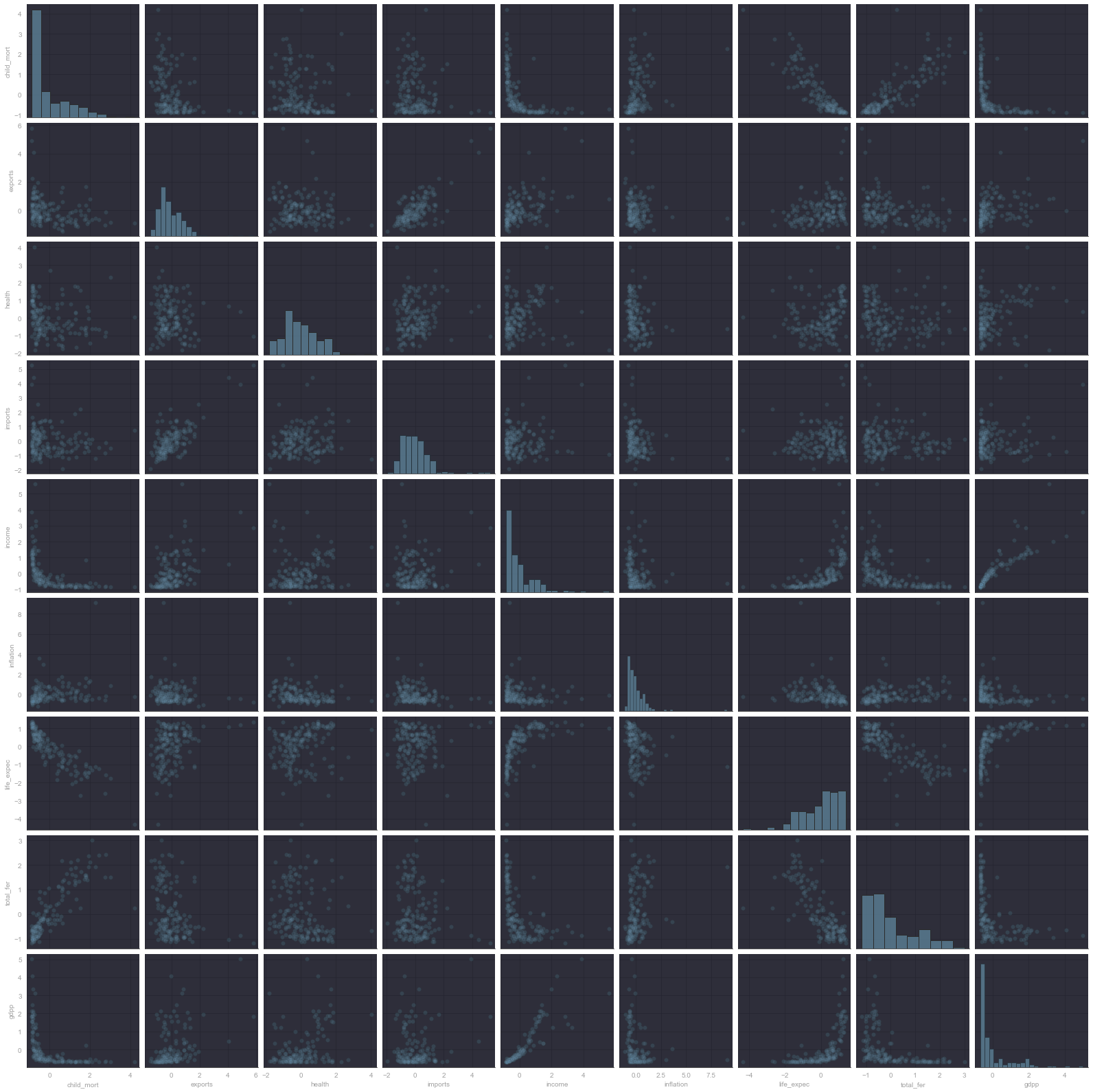


# بخش سوم: پیادهسازی

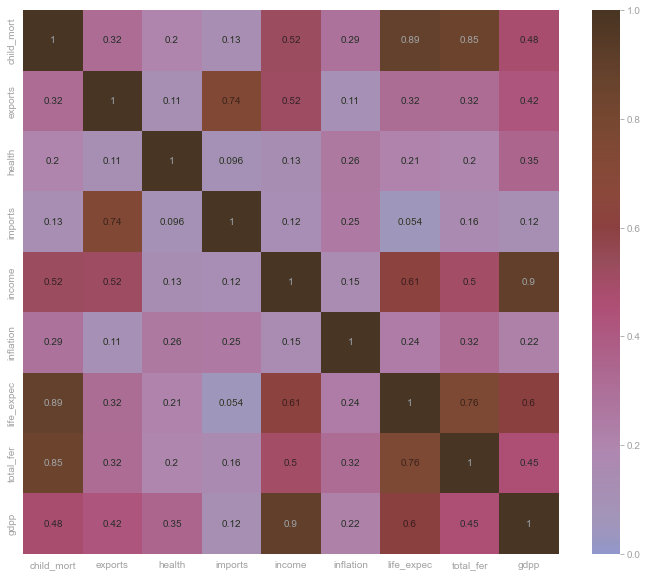
## بخش اول: داده Country

پیش پردازش داده:

ویژگی نام کشور نباید تاثیری بر نیاز آن کشور بگزارد به همین دلیل آن را حذف میکنیم. سپس نرمال سازی وارینسی میکنیم (واریانس برابر با 1). شکل کلی داده:



## همبستگی



بله، برخی از ویژگی ها همبستگی زیادی با هم دارند. برای مثال درآمد افراد و گردش ناخالص فردی، بشدت همبستگی دارند و از عکس قبلی نیز میتوان فهمید که یک شکل تقریبا خطی دارند. به همین دلیل میتوان این ویژگی ها را حذف کرد.

## نرمال سازی

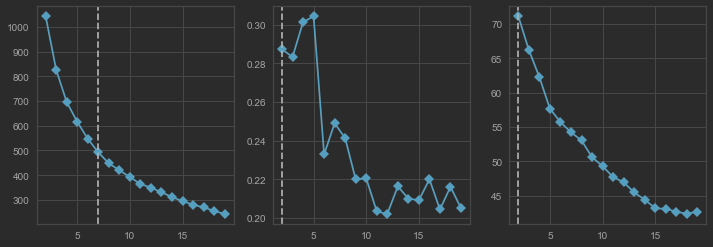
بله، زیرا که مقیاس اعداد تاثیر زیادی بر روی روش هایی مانند PCA میگزارند. برای مثال، نمره کلاس ها درس در ایران بین 0 و 20 هست اما در برخی کشور ها بین 0 تا 100 هست. اکثر روش ها در این حالت اهمیت بیشتری به عدد بزرگتر میدهند. این مشکل حتی در ویژگی های غیر مرتبط نیز وجود دارد. برای مثال ویژگی GDPP که میتواند تا بینهایت برود باید نرمال شود که بتوان اهمیت ویژگی های دیگر را نیز فهمید.

این نرمال سازی میتواند با کمک از مفاهیم ویژگی ها انجام داد و به صورت درصد نمایش داد. برای مثال تعداد مرگ ها در بین 10000 نفر را میتوان تقسیم بر 10000 کرد تا درصد بدست آید. اما این روش مشکلات خود را دارد. یک راه دیگر minmax هست که داده را بین 1 و 0 قرار میدهد (با میانگین برابر 0).

روشی که در اینجا استفاده میکنیم، std-norm هست. علت این انتخاب بجای minmax این هست برای کلاسترینگ، فواصل بین دو نقطه نشان دهنده تفاوت آن هاست و نباید فاصله نقاط پرت حداکثر 1 باشد.

## Elbow for k means

'distortion', 'silhouette', 'calinski\_harabasz'



5 یا 6 برای این مدل خوب هست.

## Silhouette explained

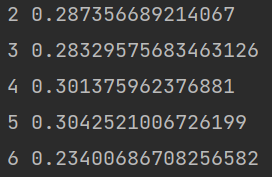
این معیار خروجی بین -1 و 1 دارد (1 یعنی عالی) و مقدار 0 به معنای خوشه های رو همدیگه دارد. به صورت کلی میانگین فواصل نقاط درون یک خوشه را با میانگین فاصله اون نقاط با نقاط نزدیکترین خوشه دیگر مقایسه میکند.

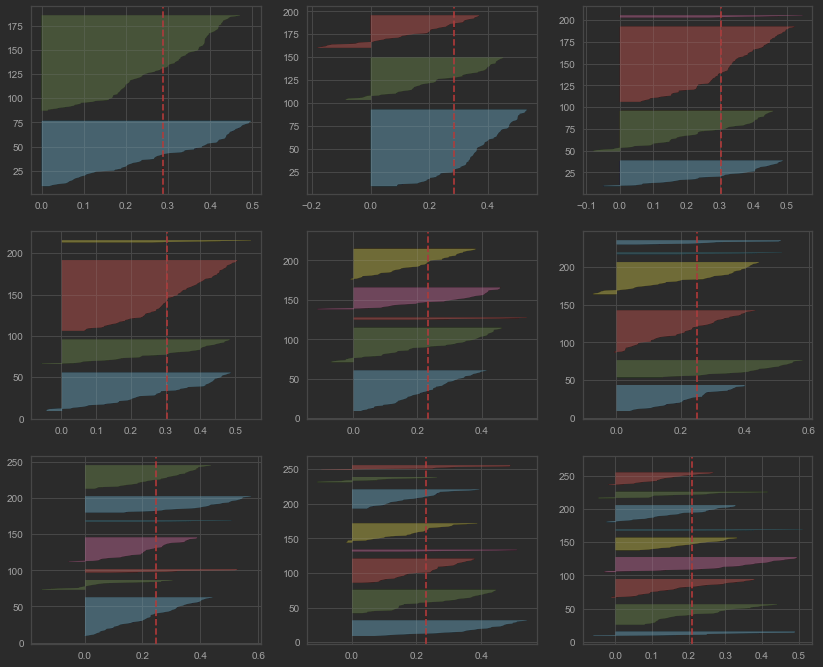
A = mean intra cluster distance

B = mean nearest cluster distance

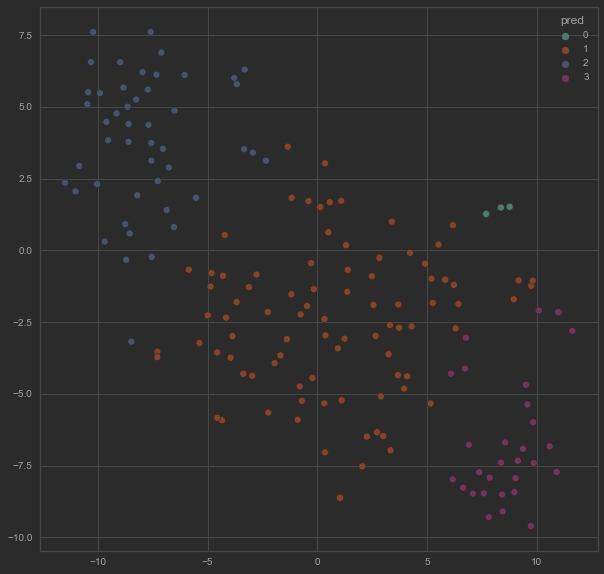
(b-a)/max(a,b)

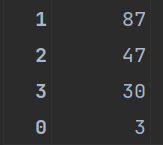
## Silhouette for k means



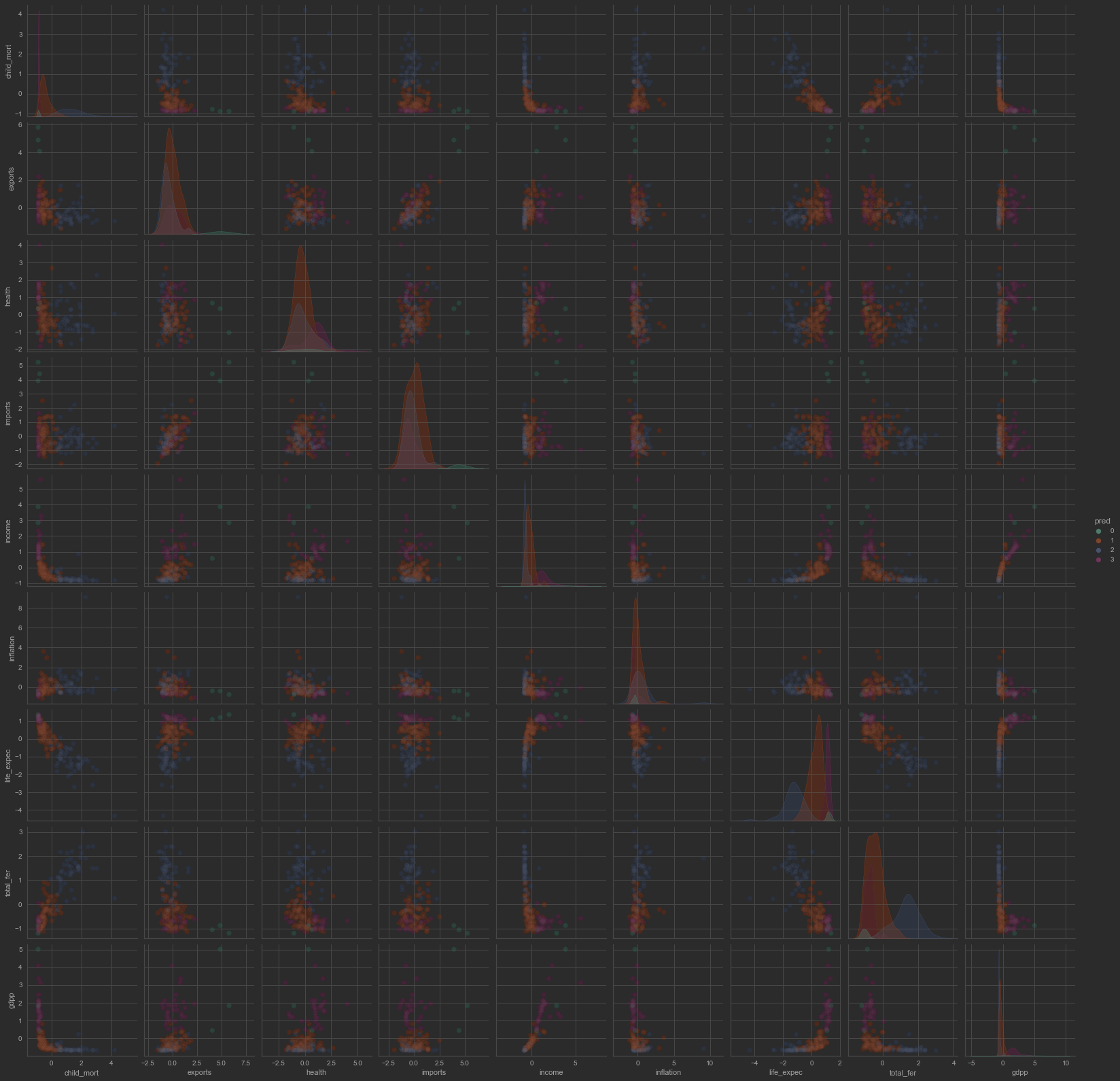


## K means for optimal k

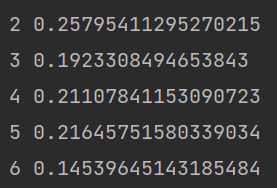
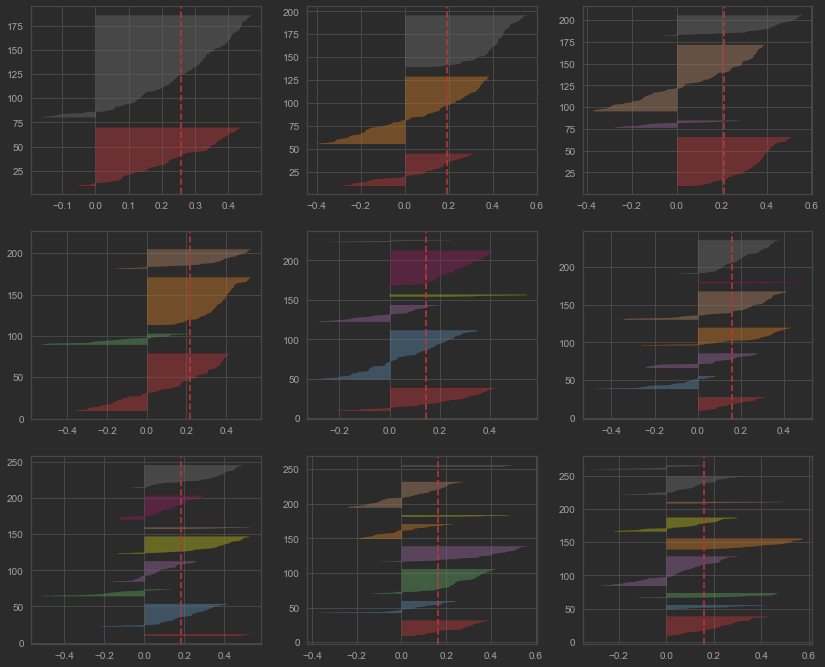
خروجی TSNE  




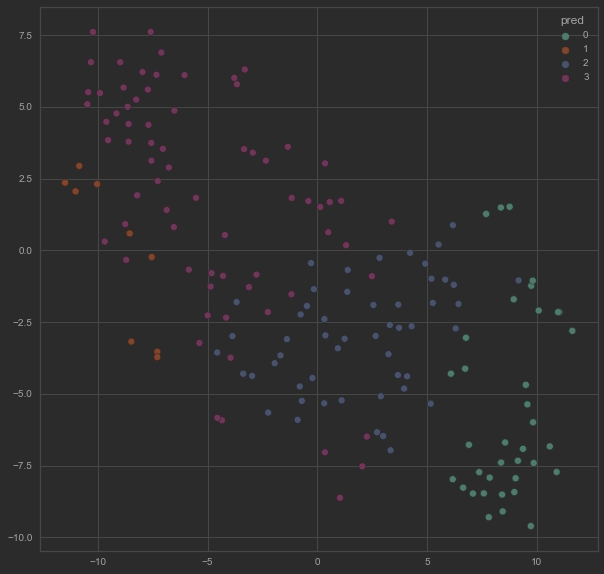
## مقایسه ویژگی ها با تصویر

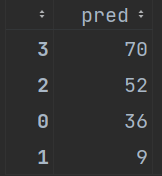


## Silhouette for GMM

## GMM with optimal k

نتیجه روی TSNE  




## مقایسه ویژگی ها با تصویر



## مقایسه GMM با Kmeans

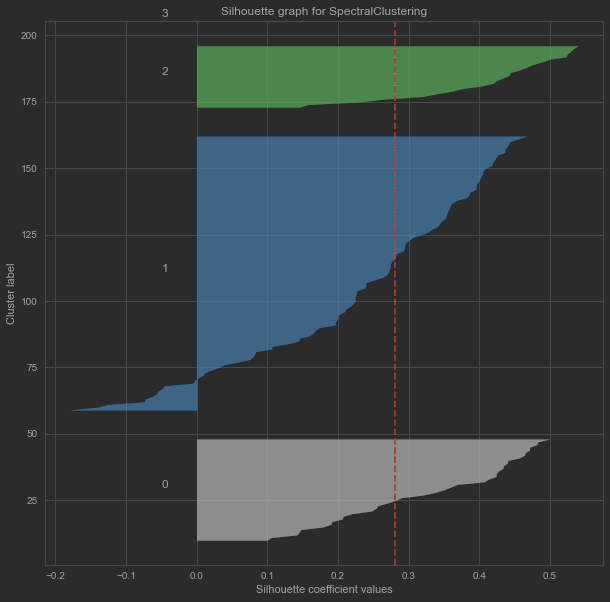
بر روی مدل Kmeans، خوشه ها بر روی TSNE به راحتی قابل تفکیک هستند. این امر بر روی GMM ضعیف تر هست. اما تعداد نقاط درون خوشه ها GMM متوارن تر هست.

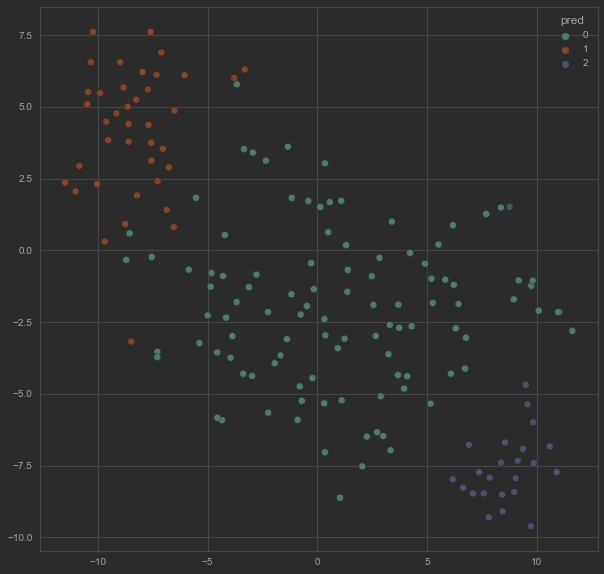
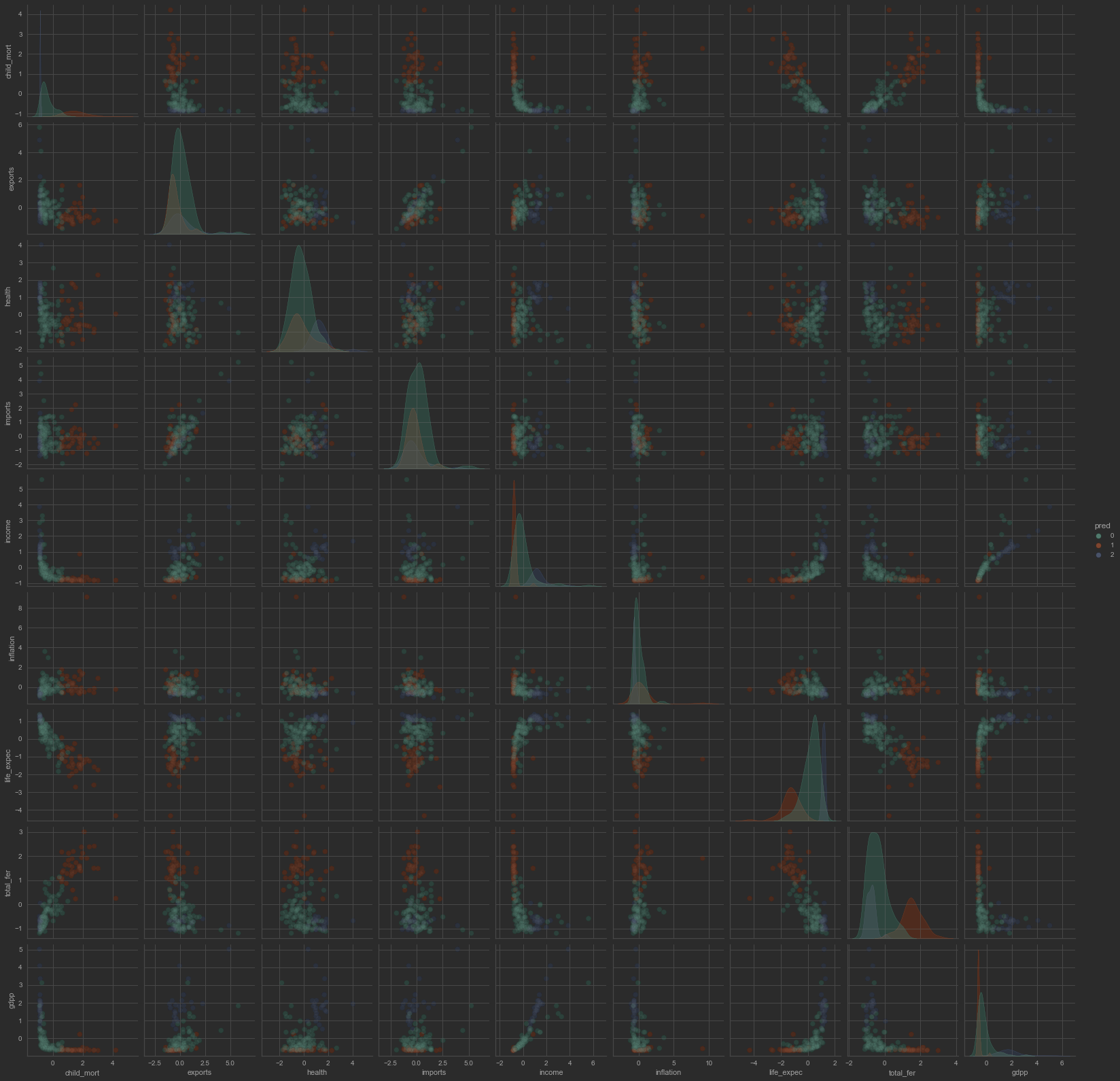
همچنین GMM کمی سریعتر اجرا میشود.

## بخش دوم: Spectral Clustering

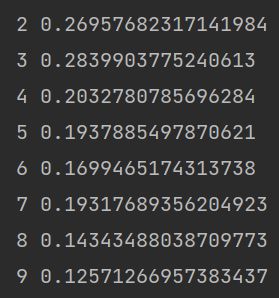
## Custom Spectral Clustering (with kneighbors)

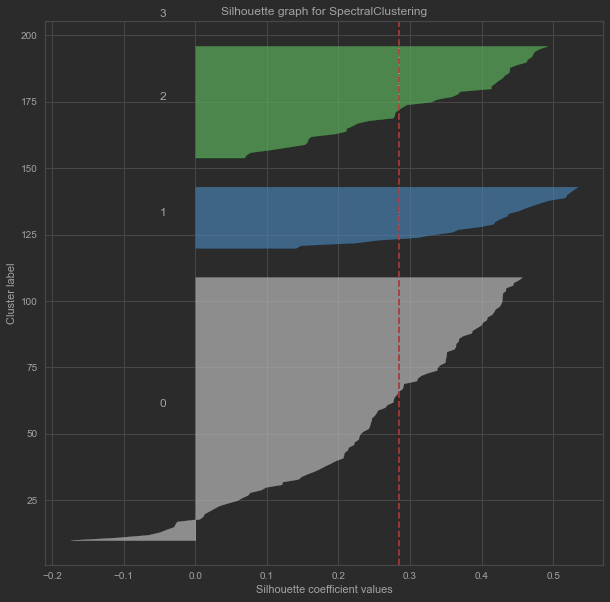


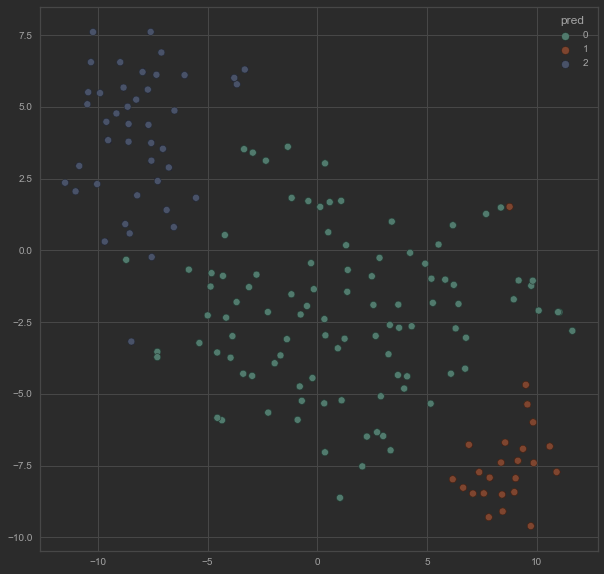
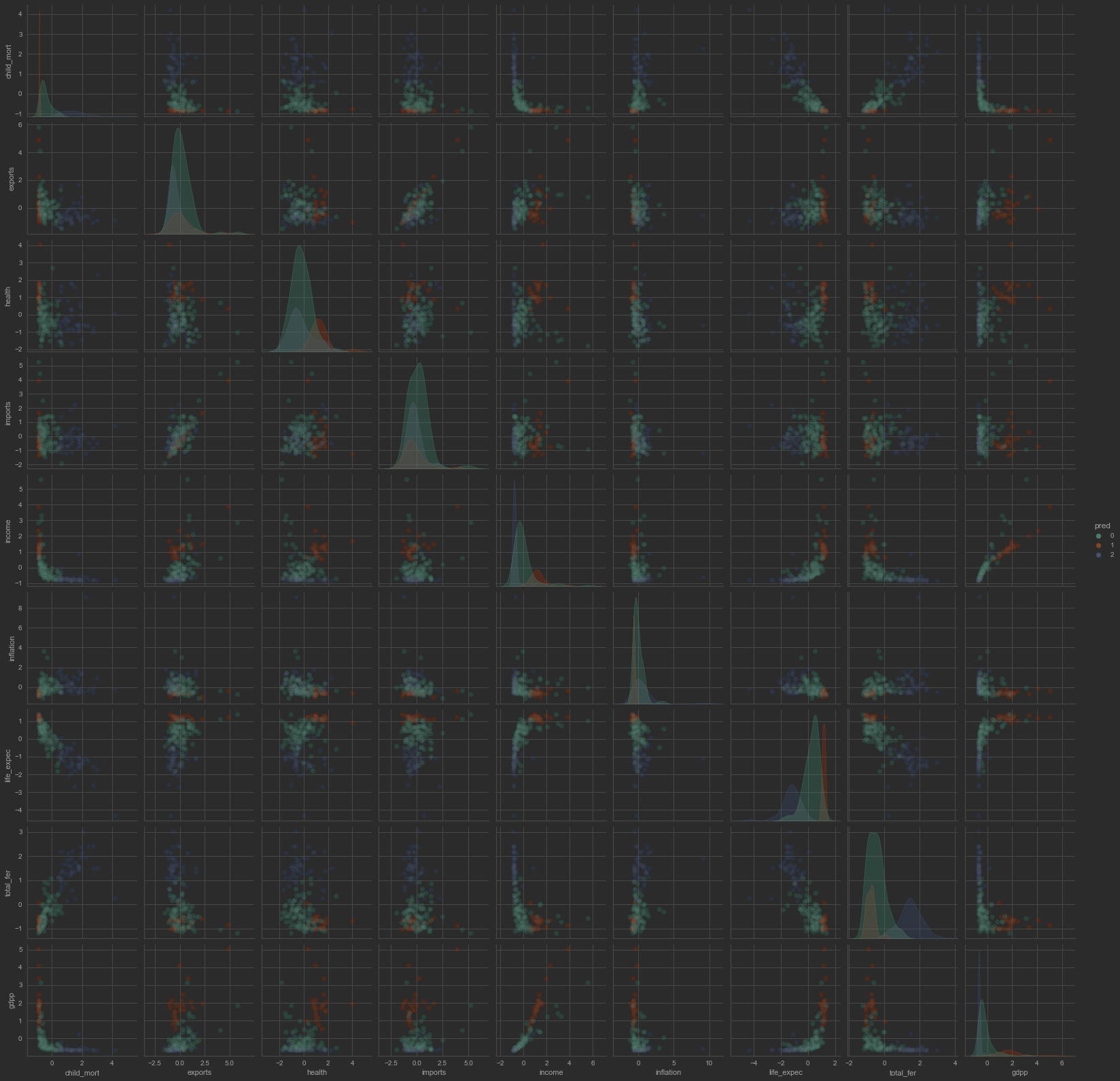


## ScikitLearn Spectral Clustering (with nearestneighbors)





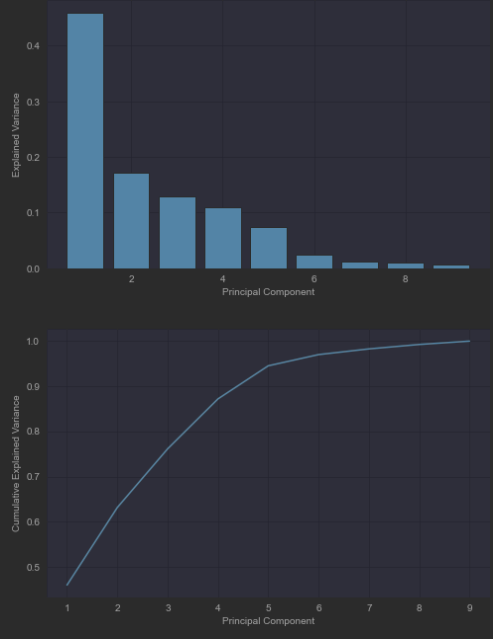
## مقایسه با kmeans و GMM

با انتخاب الگ مناسب برای تولید Affinity matrix، سرعت خوبی دارد. در صورت انتخاب الگ سنگین تر مثل RBF، بشدت کند میشود (حتی پس از RBF، استخراج بردار های ویژه نیز طول می کشد).

کیفیت این کلاستر ها نسبت به الگ های ساده تر مثل kmeans کمتر هست (silhouette score) اما با GMM قابل مقایسه هست.

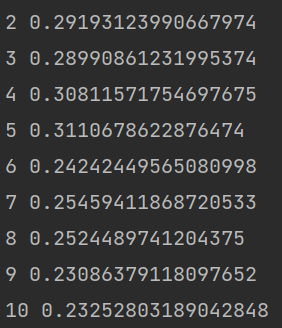
## بخش سوم: PCA

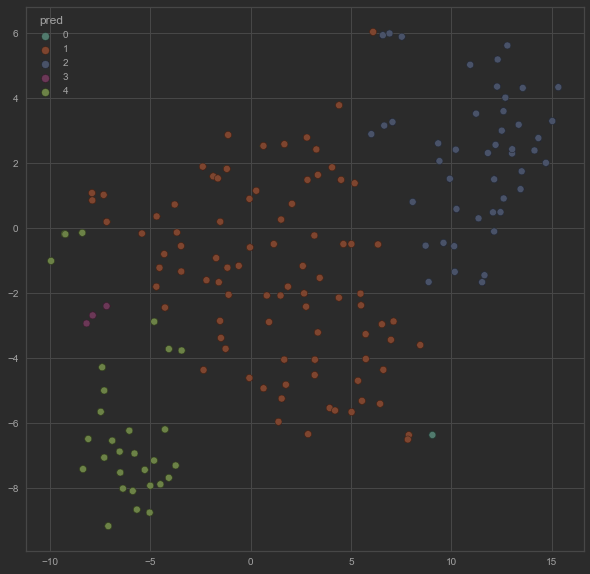
## موئلفه های مهم



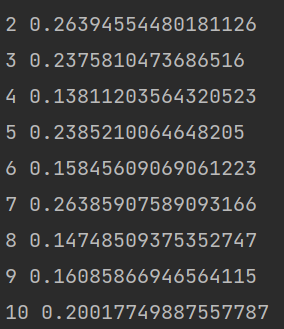
## مقایسه کلاسترینگ ها

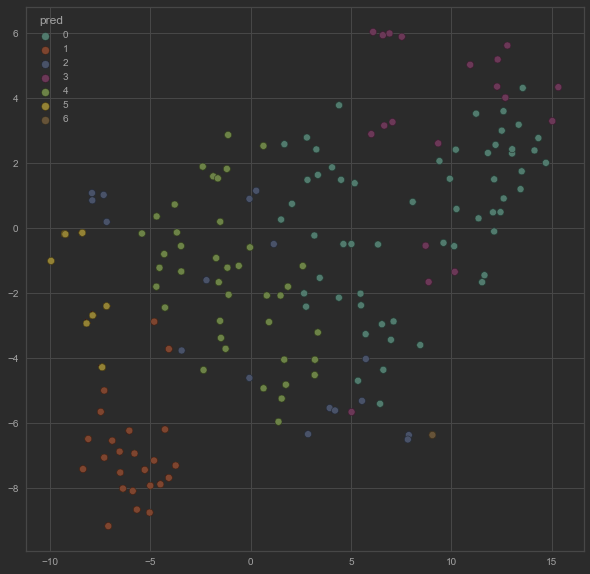
KMeans



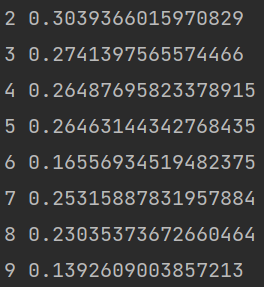


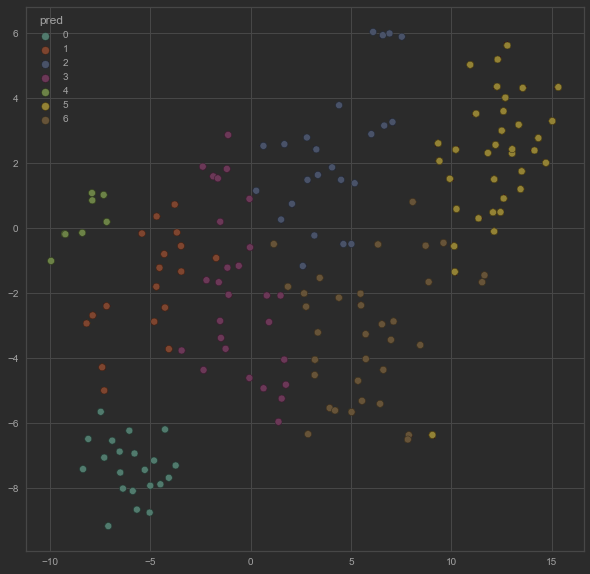
GMM





Spectral Clustering



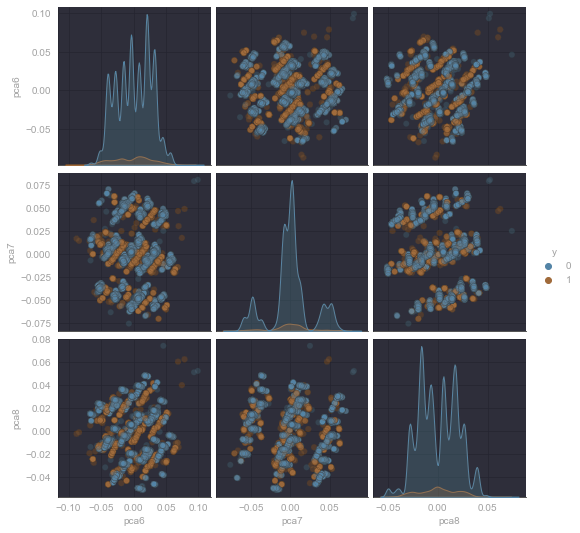


## بخش چهار: داده Shuttle و outlier

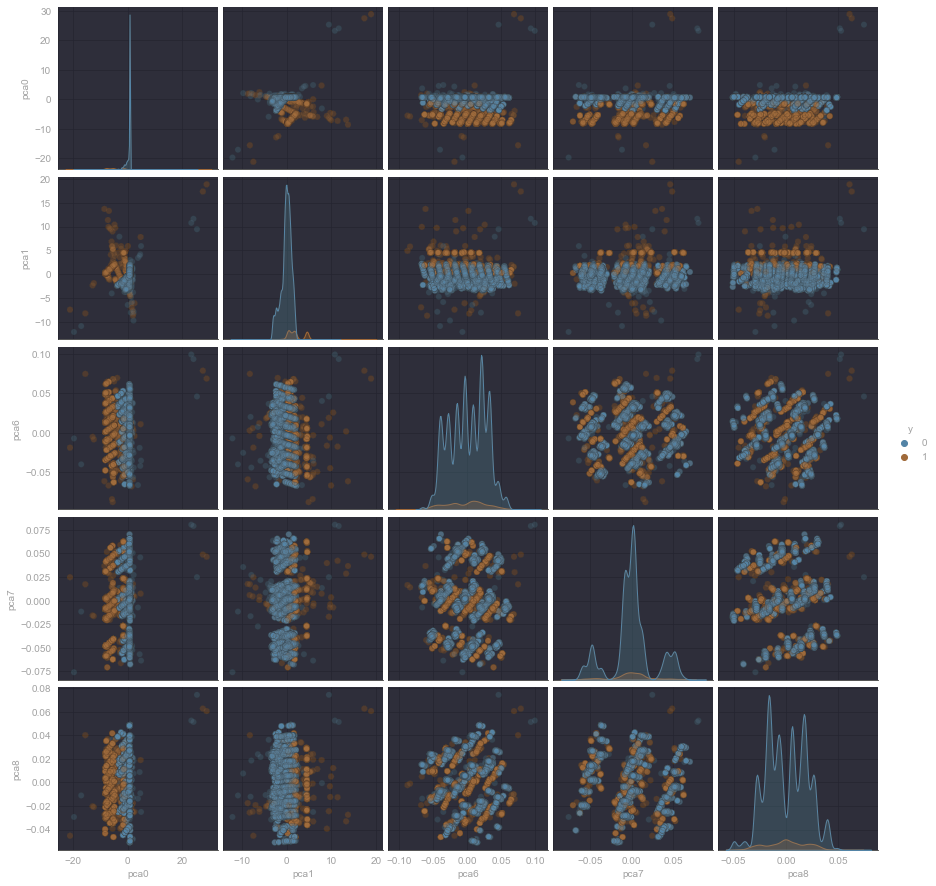
پیش پردازش داده:

نرمال سازی واریانسی به همراه PCA انجام شد.

همبستگی داده برای ستون هایی که حذف میشوند:

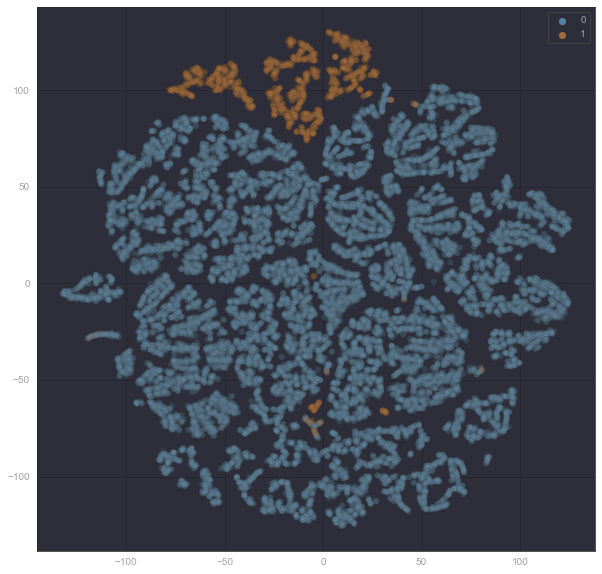


همبستگی برای دو ویزگی آخر (که حذف میشوند) و 2 ویژگی اول:

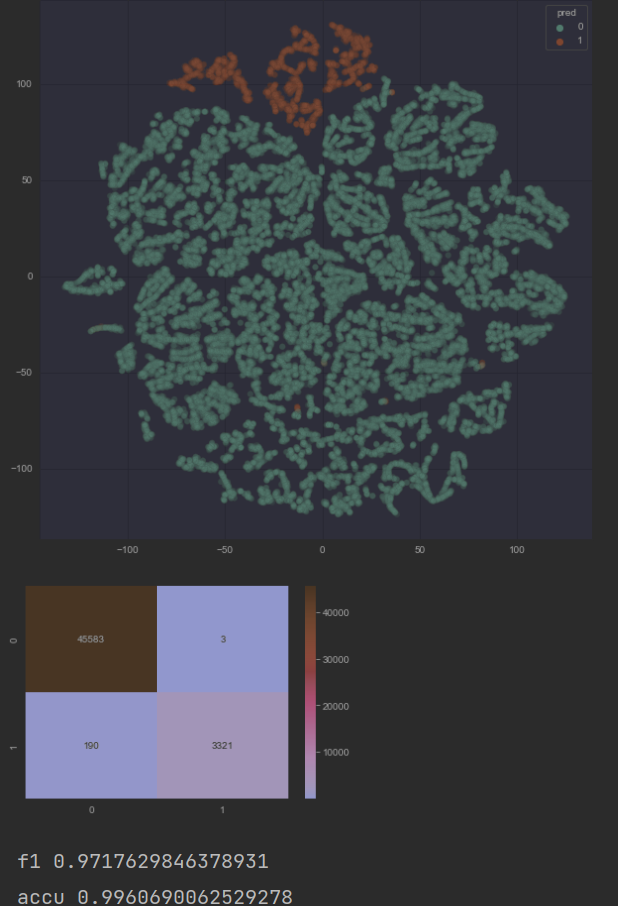


## تشخیص outlier

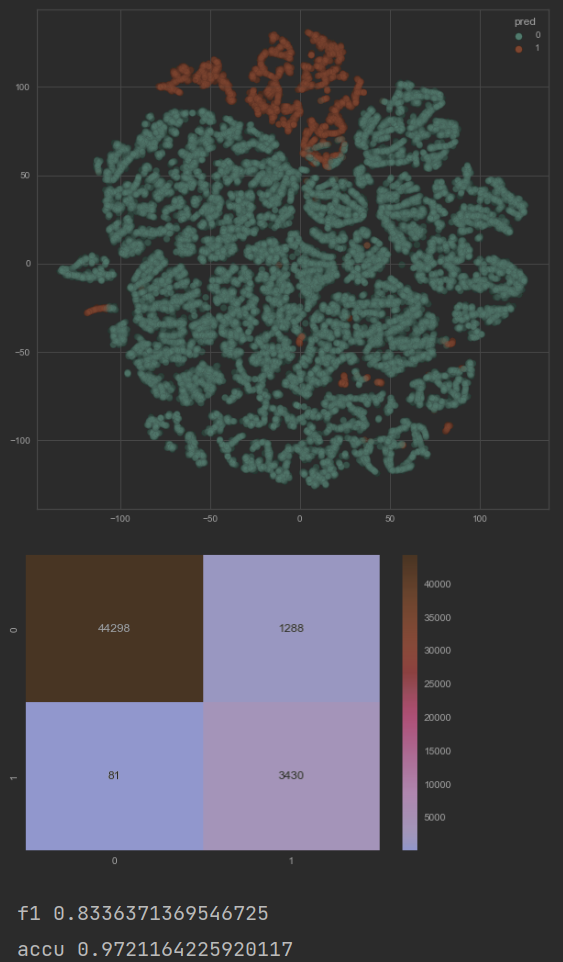
تصویر اصلی با TSNE



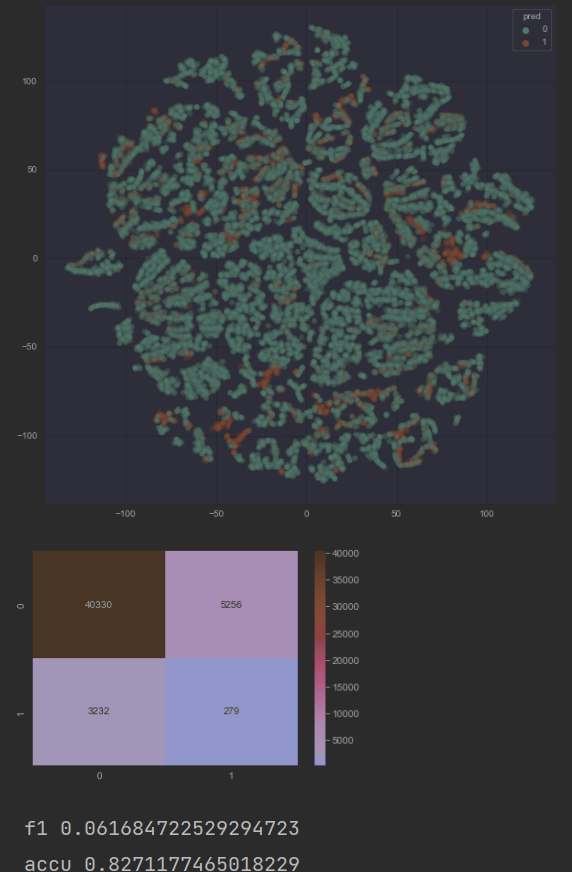
Kmeans



GMM



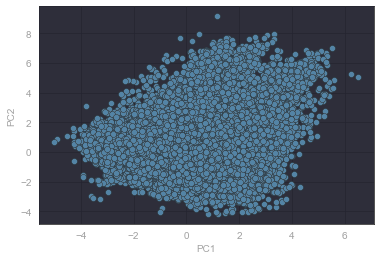
Spectral clustering (Nearest Neighbors)



## بخش پنچم: پیاده سازی دستی

داده انتخاب دارای بیش از 270 هزار داده در رابطه با شبیه سازی تقلب بانکی هست.

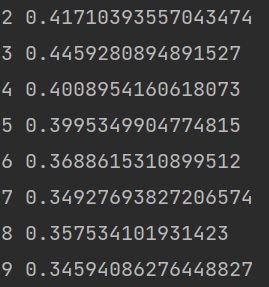
پس از پیش پردازش های مانند داده قبلی، 150 هزار تا را به PCA میدهیم و 2 ویژگی برتر را بر میداریم.

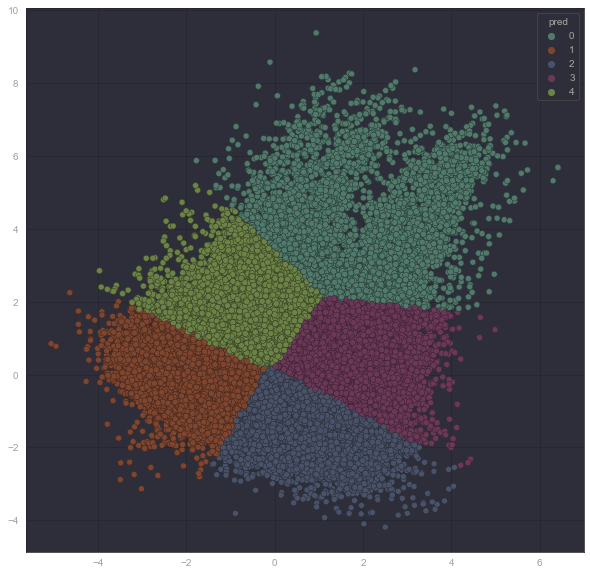


## KMeans

زمان اجرا: 1.87225

Silhouette:

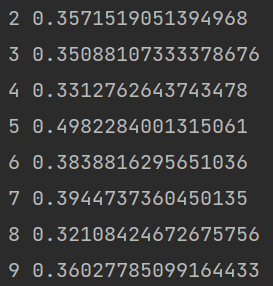


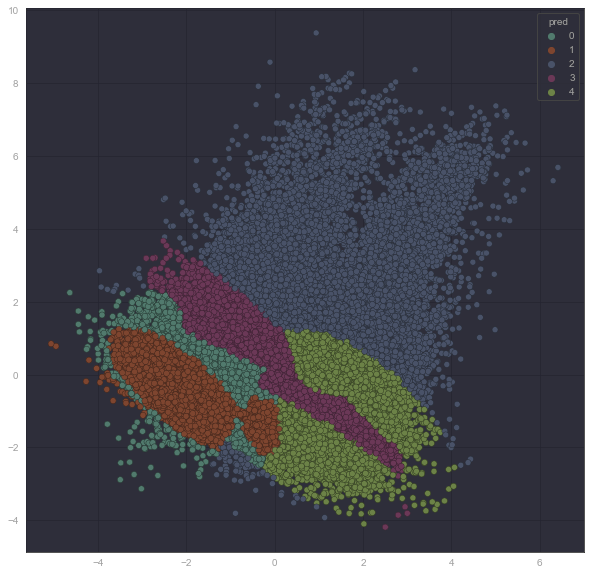


## GMM

زمان اجرا: 0.084

Silhouette:







Amirkabir University of Technology  
(Tehran Polytechnic)

Department of Computer Science and Math

Machine Learning Assignment #4

By

Seyed Hossein Mohammadi

Taught by

Dr. Akbari

dec of 2022