دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده ریاضی و علوم کامپیوتر

کارشناسی ارشد علوم کامپیوتر گرایش داده کاوی

پروژه شماره ۴ درس یادگیری ماشین

نگارش

حدیث حق‌شناس جزی

استاد راهنما

محمد اکبری

استاد مشاور

محمد علی سفیدی اصفهانی

دی ۱۴۰۱

چکیده

در این پروژه پس از فراخوانی داده ۱۶۷ کشور (شامل ۱۰ ستون از اطلاعات سلامت و درآمد و میزان امید به زندگی و ...) و انجام بررسی های ابتدایی به کمک معیار های elbow method و silhouette score از چند طریق با بررسی میزان نزدیکی این عدد به ۱ و یا تعداد اجزای منفی و ... به انتخاب تعداد خوشه بندی پرداخته ایم سپس با انواع الگوریتم های kmeans, spectral , GMM خوشه بندی را بر روی داده انجام داده و به کمک اضافه کردن لیبل هر خوشه که به داده اضافه شده است، همه ستون ها را نسبت به یکدیگر نمودار نقطه ای کشیده ایم و هر رنگ را به یک خوشه اختصاص داده ایم.

قابل ذکر است که silhouette score در ادامه توضیح داده میشود همچنین نمودار های آن نیز در کد ها قابل مشاهده است. به طور کلی هرچه مقدار این عدد به ۱ نزدیکتر باشد یعنی خوشه بندی ما دارای تراکم بیشتر و همچنین جدا پذیری بهتر است و هرچه به منفی ۱ نزدیک شود برعکس این امر اتفاق می افتد. اما باید توجه داشت که میزان این عدد از ۰ نباید کمتر شود (در این صورت خوشه بندی به درستی انجام نمیپذیرد) لذا ما به دنبال عددی برای خوشه بندی هستیم که تعداد مولفه منفی آن صفر یا کمتر از تعداد خوشه بندی دیگر باشد و همچنین عدد بالاتری نسبت به اعداد دیگر نسبت داده شده به خوشه بندی های دیگر داشته باشد.

توضیحات کلی :

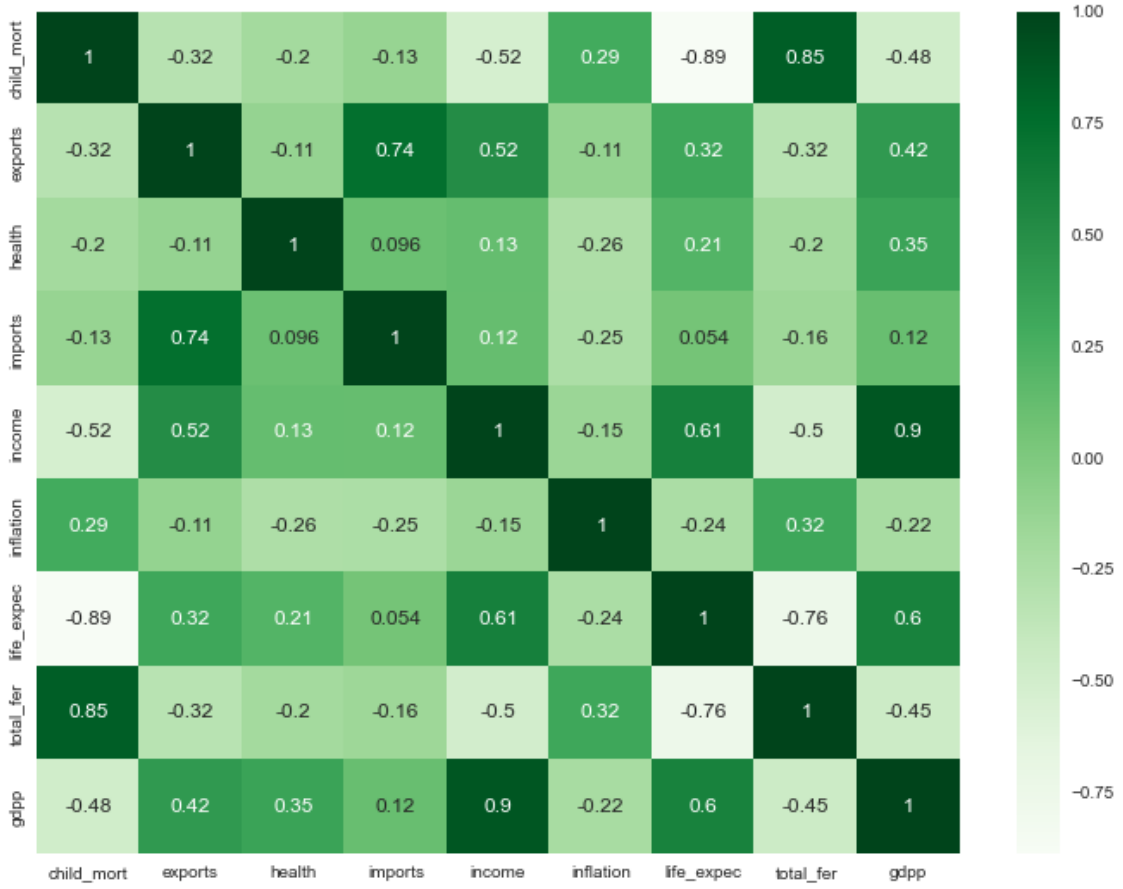
- تمرین تشریحی در یک فایل pdf موجود میباشد.
- توضیح تمامی قسمت کد ها در فایل ژوپیتر قابل مشاده است. (متاسفانه ذخیره به صورت تکست کار نکرد وامکان آوردن کد ها در این فایل نبود)
- به جای دو به دو نمودار کشیدن در بخش انتخاب ۳ مولفه، تمامی مولفه ها نسبت به یکدیگر نمودار شده اند (به نظر بهتر می آمد)
- پس از مرحله اعمال PCA و ایجاد یک دیتا فریم جدید، این دیتا فریم در فایلی به نام ml4.pca آمده و مراحل خواسته شده بر روی این داده در یک فایل جدا انجام گرفته است .

معیار silhouette score :

یکی از روش‌های ارزیابی خوشه‌بندی، معیار (Silhouette) است (که به آن معیار نیم رخ نیز گفته میشود). این معیار هم به پیوستگی درون خوشه‌ها و هم به میزان تفکیک‌پذیری آن‌ها بستگی دارد. مقدار نیم‌رخ برای هر نقطه، میزان تعلق آن را به خوشه‌اش در مقایسه با خوشه مجاور اندازه می‌گیرد. مقدار این شاخص بین ۱- تا ۱+ تغییر می‌کند. مقدار نزدیک به ۱ بیانگر انطباق خوب بین نقطه و خوشه‌اش نسبت به خوشه مجاور است. اگر معیار نیم‌رخ برای همه نقاط درون خوشه‌ها نزدیک به ۱ باشد، عمل خوشه‌بندی به درستی انجام شده است. در حالیکه کوچک بودن مقدار نیم‌رخ برای خوشه‌ها، بیانگر ضعیف بودن نتایج خوشه‌بندی است که ممکن است به علت انتخاب نامناسب تعداد خوشه‌ها (k) نیز باشد. اگر میانگین مقدار نیم‌رخ برای نقطه‌های هر خوشه را محاسبه کنیم، معیاری برای ارزیابی هر خوشه بدست می‌آید. همچنین میانگین کل مقدارهای نیم‌رخ نیز معیاری برای ارزیابی عملیات خوشه‌بندی محسوب می‌شود.

برای تفسیر این معیار، از نموداری استفاده می‌شود که میزان انطباق هر نقطه را با خوشه خودش نمایش می‌دهد. در تصویر زیر این نمودار دیده می‌شود. محور افقی نقطه‌ها و ستون‌ها، مقدار معیار نیم‌رخ برای آن نقطه است. همچنین میانگین شاخص نیم‌رخ برای همه نقاط نیز در نمودار مشخص می‌شود. اگر در این نمودار بعضی نقاط دارای مقدار نیم‌رخ منفی باشند نشان می‌دهد که ممکن است به درستی خوشه‌بندی نشده باشند و به خوشه مجاور تعلق داشته باشند.

توضیح نمودار ها:



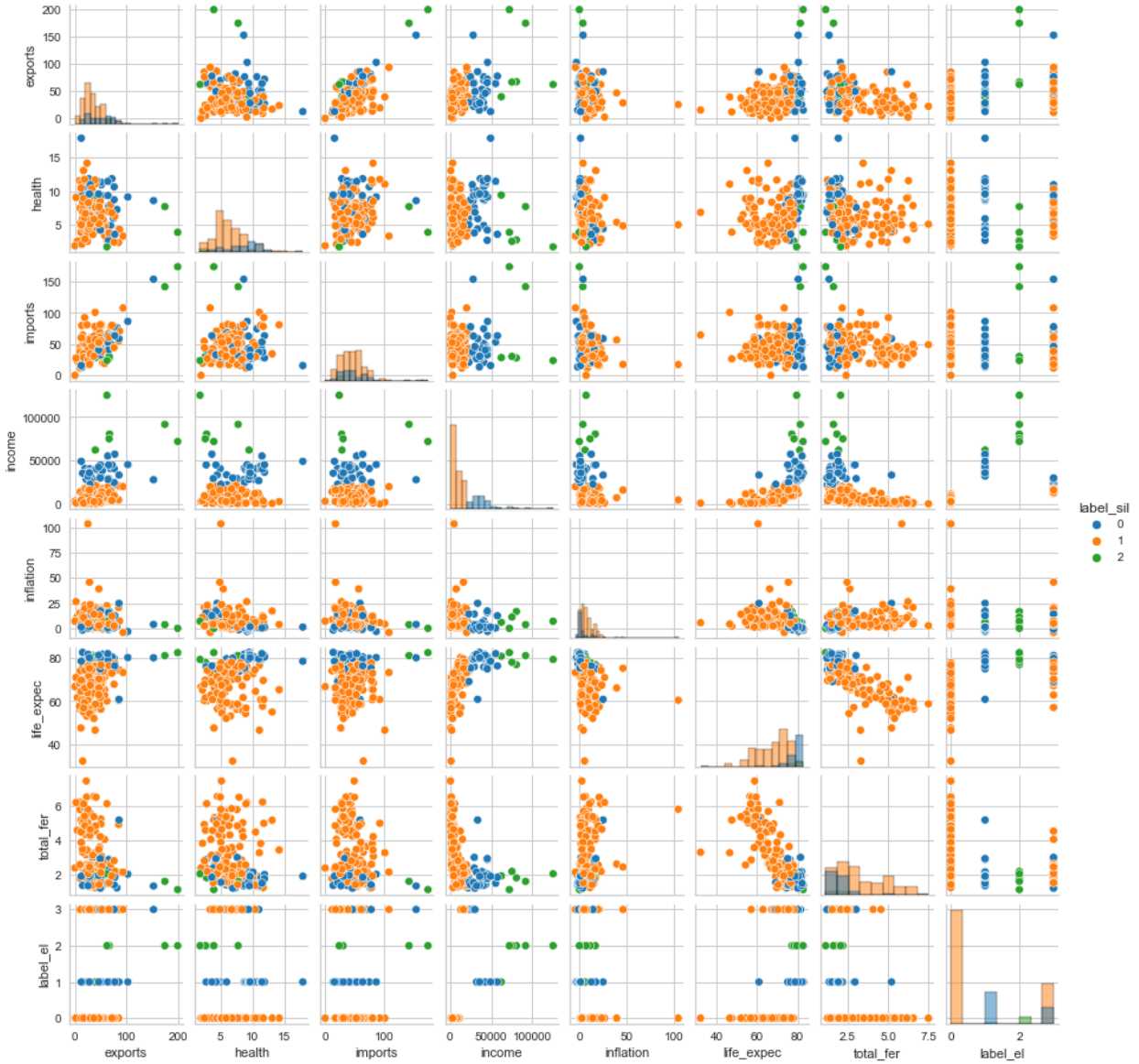
در نمودار بالا مشاهده میشود که بین ستون های gdpp و income ضریب همبستگی ۰,۹ وجود دارد . همچنین بین دو ستون child_mort و total_fer این مقدار ۰,۸۵ است. هر چه ضریب همبستگی بیشتر باشد، امکان پیش‌بینی مقدار یکی از متغیرها برحسب دیگری بیشتر است لذا دو ستون از بین این ۴ ستون حذف میکنیم چون اطلاعات نزدیک به یکدیگر در بر دارند.

خوشه بندی kmeans با تعداد خوشه elbow method:

نسبت income به life_expec : داده ها به صورت curve هستند و خوشه بندی به خوبی داده ها را از یکدیگر تمیز داده است . میتوان گفت که داده های پرت در خوشه سبز قرار گرفته اند. هرچه امید به زندگی بیشتر میشود، میزان درآمد نیز افزایش دارد مخصوصا در نقاط انتهایی که با شیب زیاد میزان امید به زندگی در کشور هایی با درآمد بالا افزایش یافته است (و بالعکس)

نسبت income به imports : همبستگی داده ها خنثی است و داده ها به خوبی خوشه بندی شده اند . خوشه سبز رنگ شامل داده های پرت میشود. با احتمال نسبتا ضعیف به میزان درامد, واردات وجود داشته است.

نسبت income به health : لزوما با افزایش درآمد کشور وضعیت سلامت بهبود نیافته است اما گروه هایی که در یک وضعیت نسبی سلامت هستند در خوشه های یکسان قرار گرفته اند.



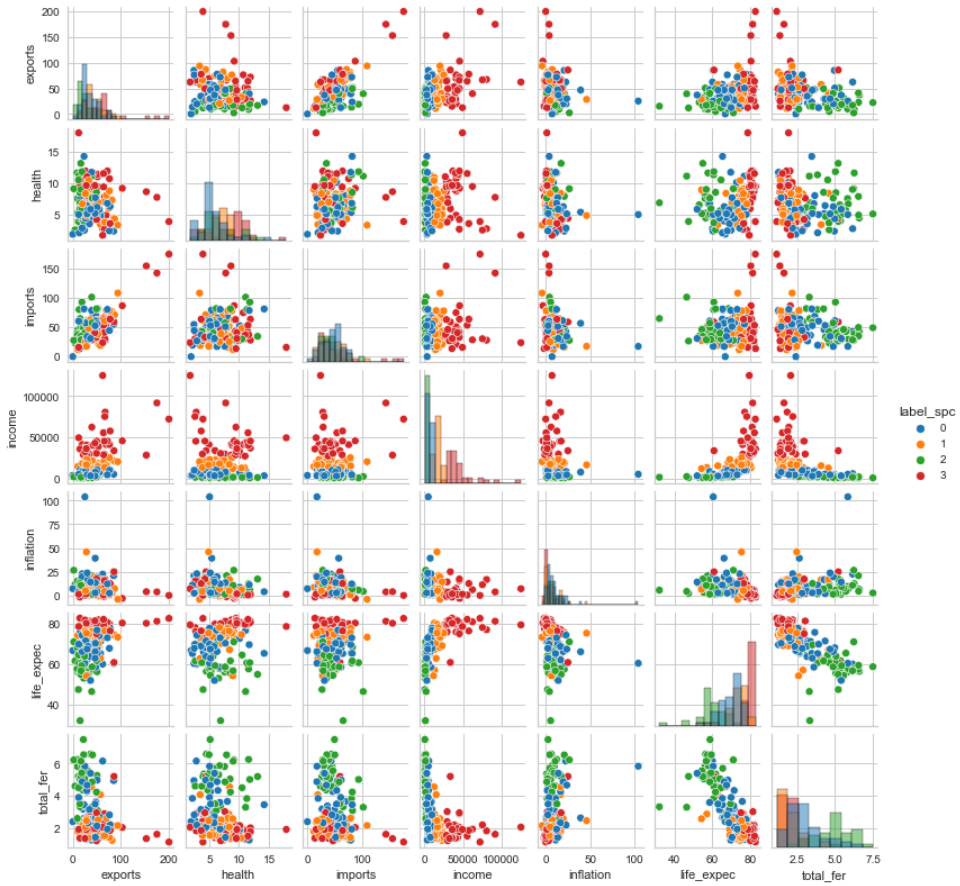
خوشه بندی kmeans با silhouette score :

نسبت income به life_expec : داده ها به صورت curve هستند و خوشه بندی به خوبی داده ها را از یکدیگر تمیز داده است . میتوان گفت که داده های پرت در خوشه سبز قرار گرفته اند. هرچه امید به زندگی بیشتر میشود، میزان درآمد نیز افزایش دارد مخصوصا در نقاط انتهایی که با شیب زیاد میزان امید به زندگی در کشور هایی با درآمد بالا افزایش یافته است (و بالعکس)

نسبت income به imports : همبستگی داده ها خنثی است و داده ها به خوبی خوشه بندی شده اند . خوشه سبز رنگ شامل داده های پرت میشود. با احتمال نسبتا ضعیف به میزان درامد, واردات وجود داشته است.

نسبت income به health: مانند خوشه بندی قبلی منتها به طور کلی به نظر من خوشه بندی ۴ تایی بهتر از خوشه بندی ۲ تایی عمل کرده است و elbow method عدد بهتری به ما داده است .


exports
health
imports
income
inflation
life_expec
total_fer
label_gmm
0
1
2
3

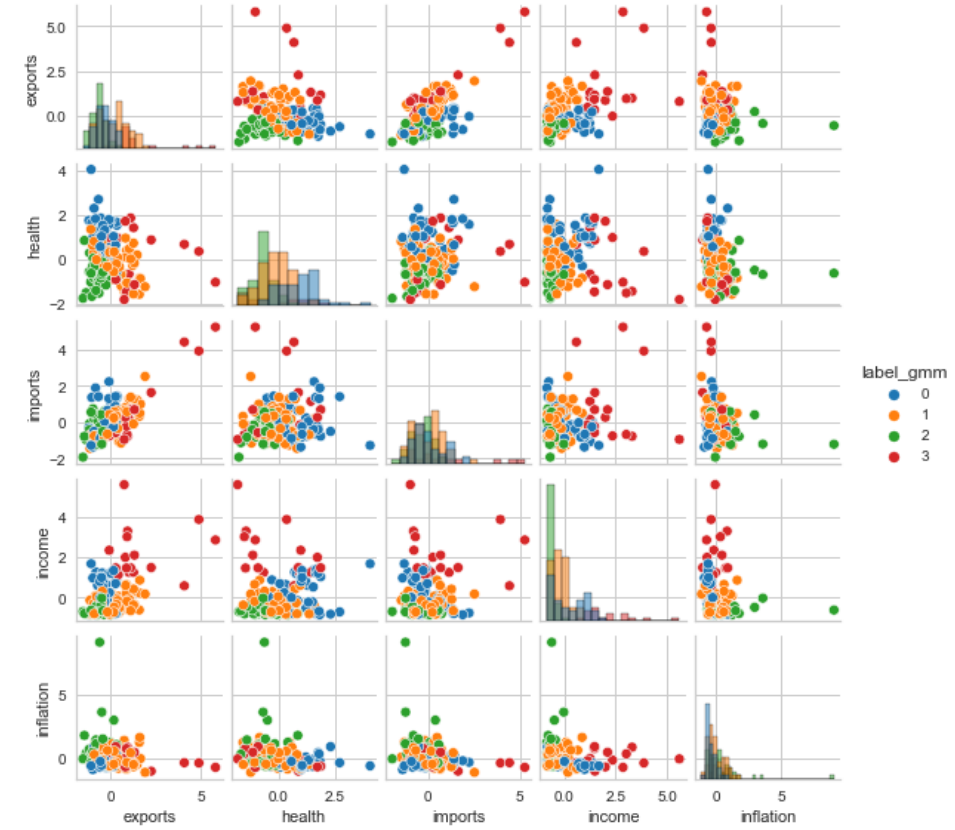

exports
health
imports
income
inflation
life_expec
total_fer
label_spc
0
1
2
3

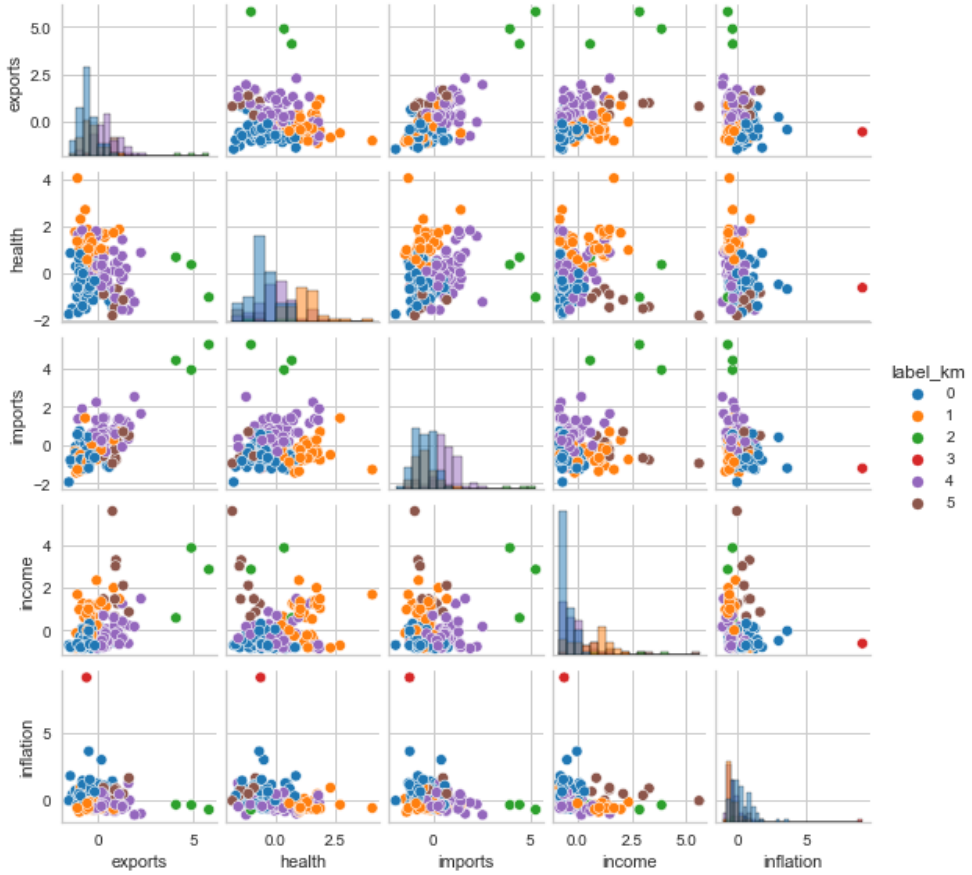
نمودار های GMM و Spectral clustering :

نسبت income به life_expec : نتایج مانند نمودار پیشین است . به نظر می آید که در GMM خوشه بندی کمی دقیق تر صورت گرفته است

نسبت income به imports : نتایج مانند نمودار پیشین است . به نظر می آید که در spectral خوشه بندی کمی دقیق تر صورت گرفته است

نسبت income به health : نتایج مانند نمودار پیشین است . به نظر می آید که در spectralخوشه بندی کمی دقیق تر صورت گرفته است


exports
health
imports
income
inflation
label_gmm
0
1
2
3

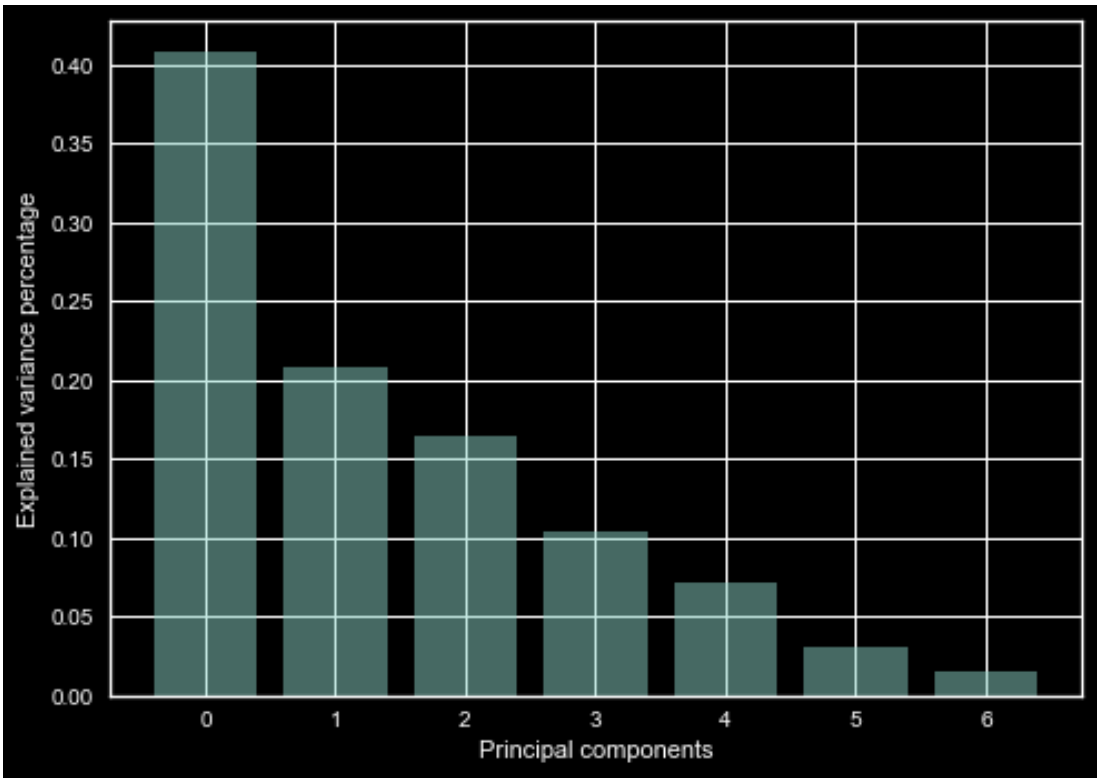

exports
health
imports
income
inflation
label_km
0
1
2
3
4
5

به ترتیب بعد از اعمال pca سه نمودار از دسته های kmean و GMM و spectral clustering :

نسبت income به infilation : داده ها به صورت خطی هستند و خوشه بندی کیفیت نسبتا پایین تری دارد . هرچه میزان درآمد کشور بالا رفته است تاثیری بر inflation نداشته است. همچنین در خوشه بندی spectral بهتر از دوتای دیگر خوشه بندی صورت پذیرفته است.

نسبت health به imports : نسبتا هرچه میزان سلامت بالاتر رفته است میزان import هم افزایش داشته است و بهترین خوشه بندی به کمک الگوریتم GMM انجام شده است.

نسبت health به income : بهترین خوشه بندی توسط الگوریتم های spectral و kmeans صورت گرفته است. در خوشه بندی آخر میتوان به خوبی مشاهده کرد که داده های پرت در یک خوشه قرار گرفته اند .



(نمودار مربوط به pca . همانطور که مشاهده میشود ستون های اول تا پنجم بیش از ۹۰٪ تاثیر گذاری و پراکندگی را در دیتا دارند پس حفظ میشوند و مابقی را حذف میکنیم.

پایان