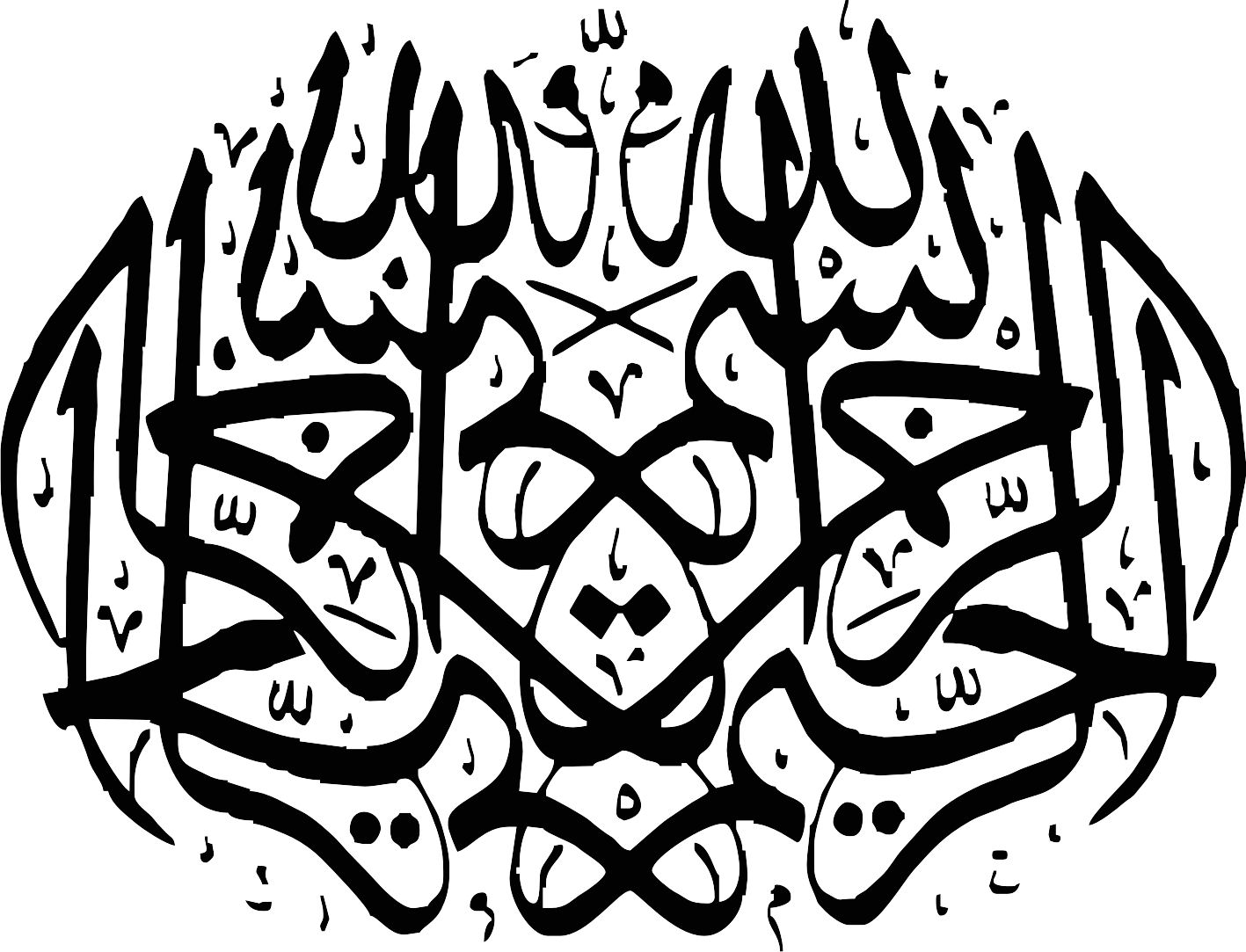


**گزارش تکلیف پنجم درس یادگیری ماشین کاربردی**

**استاد درس: دکتر ناظرفرد**

تهیه کننده: سید نیما محمودیان

شماره دانشجویی: 402125005



فهرست مطالب

[2-پاسخ سوال دوم 1](#_Toc168045514)

[1-2-آماده‌سازی داده‌ها 1](#_Toc168045515)

[2-2-مدیریت مقادیر گمشده 4](#_Toc168045516)

[3-2-آموزش مدل ماشین بردار پشتیبان 6](#_Toc168045517)

[4-2-بهینه‌سازی هایپرپارامترها 7](#_Toc168045518)

[3-پاسخ سوال سوم 9](#_Toc168045519)

[4-پاسخ سوال چهارم 12](#_Toc168045520)

[1-4-آموزش جنگل تصادفی 13](#_Toc168045521)

[2-4-آموزش سه الگوریتم یادگیری جمعی دیگر 16](#_Toc168045522)

[1-2-4-استفاده از تقویت گرادیان 16](#_Toc168045523)

[2-2-4-استفاده از آدابوست 17](#_Toc168045524)

[3-2-4-استفاده از XGBoost 17](#_Toc168045525)

فهرست شکل‌ها

[شکل 1: وارد کردن کتابخانه‌های مورد نیاز 1](#_Toc168045533)

[شکل 2: بارگذاری داده‌ها 1](#_Toc168045534)

[شکل 3: بررسی ساختار اولیه داده‌ها 1](#_Toc168045535)

[شکل 4: خروجی متد .info() 2](#_Toc168045536)

[شکل 5: تعداد مقادیر گمشده هر ستون 2](#_Toc168045537)

[شکل 6: حذف ستون‌های غیر ضروری 2](#_Toc168045538)

[شکل 7: دسته بندی ستون‌های عددی و دسته‌ای 3](#_Toc168045539)

[شکل 8: نگاشت مقادیر دسته‌ای در ستون‌های Sex و Survived 3](#_Toc168045540)

[شکل 9: OneHotEncoding 4](#_Toc168045541)

[شکل 10: حذف داده‌های پرت با استفاده از روش دامنه میان چارکی 4](#_Toc168045542)

[شکل 11: پایپ‌لاین پیش پردازش داده‌های عددی 5](#_Toc168045543)

[شکل 12: نحوه تعریف کلاس ColumnTransformer() 5](#_Toc168045544)

[شکل 13: تعریف مدل و پایپ‌لاین اصلی 6](#_Toc168045545)

[شکل 14: جایگذاری مقادیر خالی با میانه 6](#_Toc168045546)

[شکل 15: پایپ‌لاین آموزش SVM 7](#_Toc168045547)

[شکل 16: محاسبه و نمایش مقادیر معیارهای ارزیابی 7](#_Toc168045548)

[شکل 17: پارامترهای مدل 8](#_Toc168045549)

[شکل 18: انجام جست و جوی شبکه‌ای 8](#_Toc168045550)

[شکل 19: هایپرپارامترهای بهینه 8](#_Toc168045551)

[شکل 20: پایپ‌لاین با هایپرپارامترهای بهینه 8](#_Toc168045552)

[شکل 21: نتایج بهینه‌سازی پارامترها 9](#_Toc168045553)

[شکل 22: ایمپورت کردن کتابخانه‌های مورد نیاز 9](#_Toc168045554)

[شکل 23: بررسی تعداد هر یک از کلاس‌ها 9](#_Toc168045555)

[شکل 24: حذف ردیف‌های تکراری 10](#_Toc168045556)

[شکل 25: انکود کردن ستون هدف 10](#_Toc168045557)

[شکل 26: ارزیابی دقت با آرایه‌ای متشکل از صفر 10](#_Toc168045558)

[شکل 27:گزارش عملکرد در حالت پایه 11](#_Toc168045559)

[شکل 28: آموزش مدل با داده‌های بالانس شده 11](#_Toc168045560)

[شکل 29: تست مدل و ارزیابی دقت 11](#_Toc168045561)

[شکل 30: گزارش عملکرد مدل آموزش داده شده 12](#_Toc168045562)

[شکل 31: کتابخانه‌های مورد نیاز 12](#_Toc168045563)

[شکل 32: تعریف پایپ‌لاین برای آموزش یا تست جنگل تصادفی 13](#_Toc168045564)

[شکل 33: تعریف پارامترهای گرید سرچ 13](#_Toc168045565)

[شکل 34: استفاده از گرید سرچ برای بهینه‌سازی هایپرپارامترها 14](#_Toc168045566)

[شکل 35: لیست پارامترهای گرید سرچ 14](#_Toc168045567)

[شکل 36: نمودار میله‌ای به دست آمده از گرید سرچ 15](#_Toc168045568)

[شکل 37: هایپرپارامترهای بهینه و دقت به دست آمده از آن 15](#_Toc168045569)

[شکل 38: پایپ‌لاین به روز شده برای تقویت گرادیان 16](#_Toc168045570)

[شکل 39: دقت محاسبه شده از تقویت گرادیان 16](#_Toc168045571)

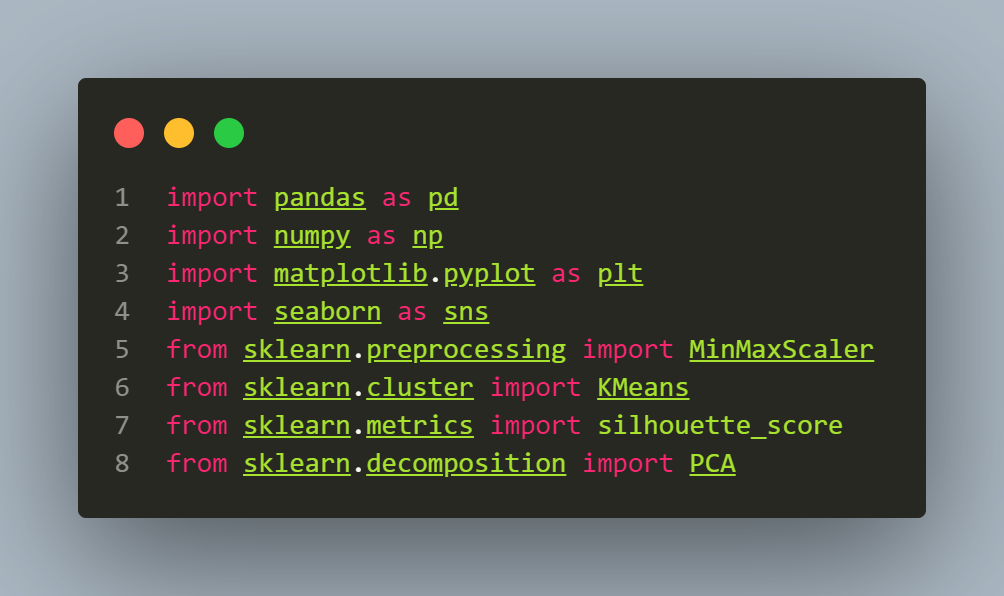
[شکل 40: فرایند آموزش و تست آدابوست 17](#_Toc168045572)

[شکل 41: فرایند آموزش و تست XGBoost 18](#_Toc168045573)

# 1-پاسخ سوال اول

## 1-1-آماده‌سازی داده‌ها

کد شکل (1)، کتابخانه‌های مورد نیاز برای پردازش و تحلیل داده‌ها، از جمله pandas، numpy و ابزارهای مختلف از scikit-learn را وارد می‌کند.



شکل 1: وارد کردن کتابخانه‌های مورد نیاز

داده‌ها بارگذاری شده و در DataFrame به نام df ذخیره می‌شوند. (شکل 2)



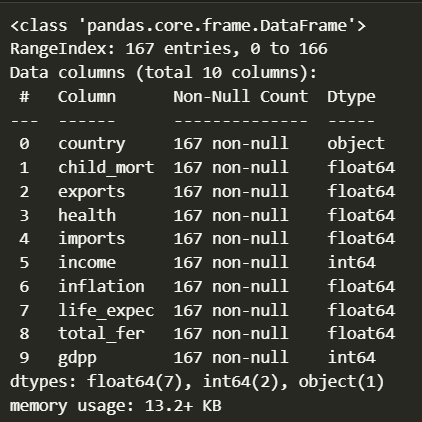
شکل 2: بارگذاری داده‌ها

سپس ساختار داده‌ها و اطلاعات اولیه آن با استفاده از ()df.info نمایش داده می‌شود تا نوع و تعداد مقادیر گمشده بررسی شود. (شکل 3)



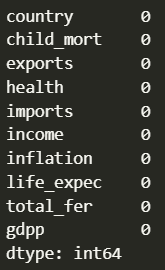
شکل 3: بررسی ساختار اولیه داده‌ها

شکل (4) خروجی این متد را نمایش می‌دهد.



شکل 4: خروجی متد .info()

سپس تعداد مقادیر گمشده در هر ستون داده‌ها با استفاده از ()df.isna().sum محاسبه می‌شود. شکل (5) خروجی این متد را نمایش می‌دهد.

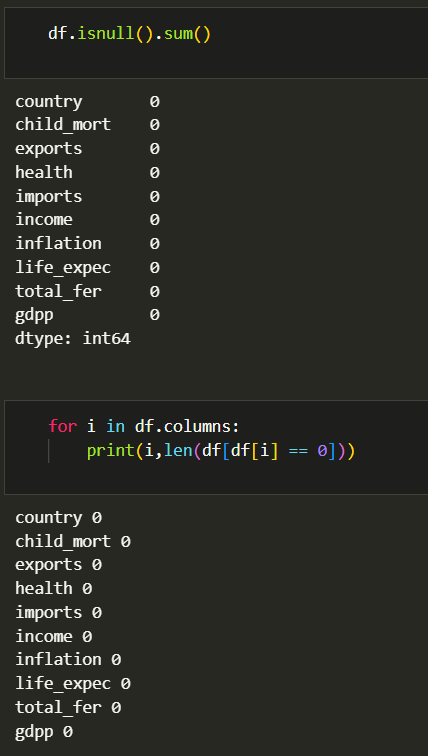


شکل 5: تعداد مقادیر گمشده هر ستون

در ادامه تعداد مقادیر صفر و "؟" در هر ستون محاسبه و نمایش داده می‌شود. مشاهده می‌شود که در هیچ ستونی مقدار صفر یا ؟ نداریم. شکل (6) کد مربوط به این کار و خروجی را نمایش می‌دهد.

## 2-1-رسم ماتریس همبستگی

ابتدا ستون 'country' را از دیتافریم حذف می‌کنیم. دلیل این کار این است که ستون 'country' برای تحلیل همبستگی مورد نیاز نیست چون شامل داده‌های اسمی است که برای محاسبه همبستگی مناسب نیستند. سپس ماتریس همبستگی را محاسبه می‌کنیم. این ماتریس به ما نشان می‌دهد که هر دو ستون عددی موجود در دیتافریم چه میزان همبستگی با یکدیگر دارند. در نهایت، ماتریس همبستگی محاسبه شده را نمایش می‌دهیم تا بتوانیم تحلیل کنیم که کدام متغیرها با یکدیگر همبستگی دارند و میزان این همبستگی چقدر است. شکل (7) نحوه انجام این کار را نمایش می‌دهد.



شکل 6: پیدا کردن مقادیر 0 و ؟



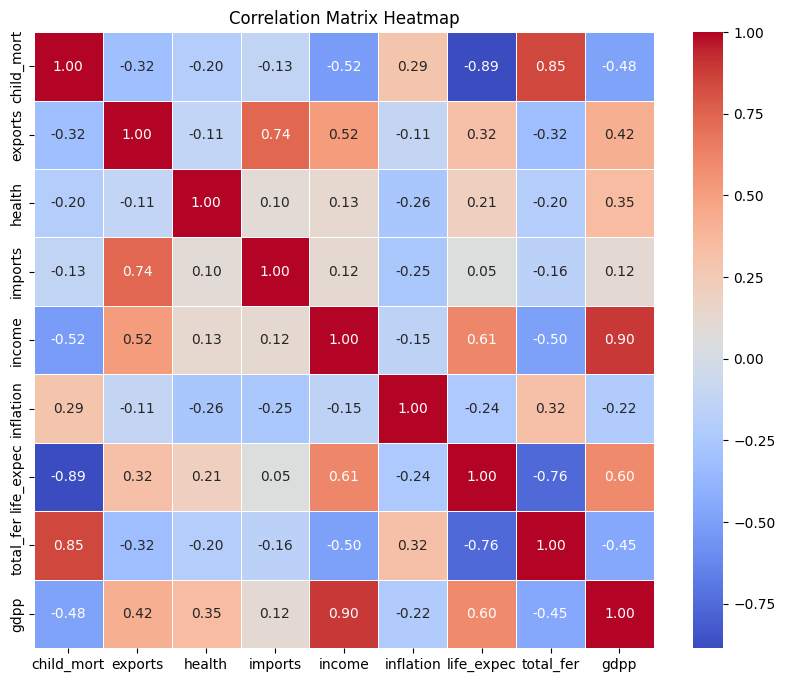
شکل 7: محاسبه ماتریس همبستگی

سپس به رسم ماتریس همبستگی می‌پردازیم. ابتدا اندازه نمودار را تنظیم می‌کنیم تا نمودار به اندازه دلخواه ما درآید. سپس یک هیت‌مپ (Heatmap) ایجاد می‌کنیم تا ماتریس همبستگی را بصری کنیم. در این مرحله، پارامترهای زیر را تنظیم می‌کنیم: annot=True: برای نمایش اعداد همبستگی روی هیت‌مپ. fmt=".2f": برای نمایش اعداد با دقت دو رقم اعشار. cmap='coolwarm': برای تنظیم رنگ‌های هیت‌مپ. linewidths=0.5: برای تنظیم ضخامت خطوط جداکننده سلول‌ها. سپس عنوان نمودار را تنظیم می‌کنیم تا بفهمیم این نمودار نمایش‌دهنده چه چیزی است. در نهایت، نمودار را نمایش می‌دهیم تا بتوانیم تحلیل بصری از ماتریس همبستگی داشته باشیم. شکل (8) نحوه انجام این کار را نمایش می‌دهد.

همچنین شکل (9) نمودار به دست آمده را نمایش می‌دهد.



شکل 8: نحوه رسم هیت‌مپ



شکل 9: رسم هیت‌مپ

ماتریس همبستگی روابط بین ویژگی‌های مختلف را نشان می‌دهد. در اینجا برخی مشاهدات آورده شده است:

* همبستگی‌های بالا:

مرگ و میر کودکان و امید به زندگی: همبستگی منفی بسیار قوی (-0.886). این نشان می‌دهد که نرخ‌های بالاتر مرگ و میر کودکان با امید به زندگی پایین‌تر مرتبط است.

مرگ و میر کودکان و باروری کل: همبستگی مثبت قوی (0.848). نرخ‌های بالاتر مرگ و میر کودکان با نرخ‌های بالاتر باروری مرتبط است.

درآمد و تولید ناخالص داخلی سرانه (GDP per capita): همبستگی مثبت بسیار قوی (0.896). درآمد بالاتر با تولید ناخالص داخلی سرانه بالاتر به شدت مرتبط است.

* همبستگی‌های متوسط:

درآمد و امید به زندگی: همبستگی مثبت متوسط (0.612). درآمد بالاتر تمایل به همراهی با امید به زندگی طولانی‌تر دارد.

صادرات و واردات: همبستگی مثبت قوی (0.737). کشورهایی که بیشتر صادر می‌کنند، تمایل به واردات بیشتر نیز دارند.

امید به زندگی و تولید ناخالص داخلی سرانه: همبستگی مثبت متوسط (0.600).

ویژگی‌هایی با همبستگی‌های پایین‌تر:

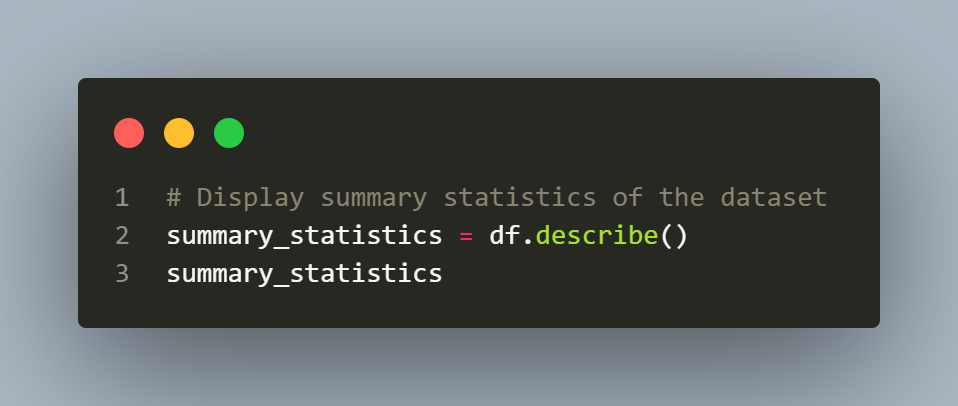
هزینه‌های بهداشتی: همبستگی‌های نسبتاً پایین‌تری با سایر ویژگی‌ها دارد، بالاترین همبستگی با تولید ناخالص داخلی سرانه (0.346) است.

واردات و امید به زندگی: همبستگی مثبت بسیار پایین (0.054).

با توجه به همبستگی‌های بالا، دو ستون "gdpp" , "child\_mort" را با متد .drop() حذف می‌کنیم. ستون “child\_mort” با دو ستون "total\_fer" و “life\_expec” همبستگی بالایی دارد پس حذف این ستون منطقی است.

## 3-1-بررسی احتیاج به نرمال سازی

ابتدا با استفاده از متد .describe() ابتدا آمار خلاصه را برای دیتاست محاسبه می‌کنیم. این آمار شامل معیارهای توصیفی مانند میانگین، انحراف معیار، کمینه، بیشینه و مقادیر چارکی است. سپس آمار خلاصه محاسبه‌شده را نمایش می‌دهیم تا بتوانیم اطلاعاتی جامع و مفید از وضعیت کلی داده‌ها به دست آوریم. شکل (10) نحوه انجام این کار را نمایش می‌دهد.



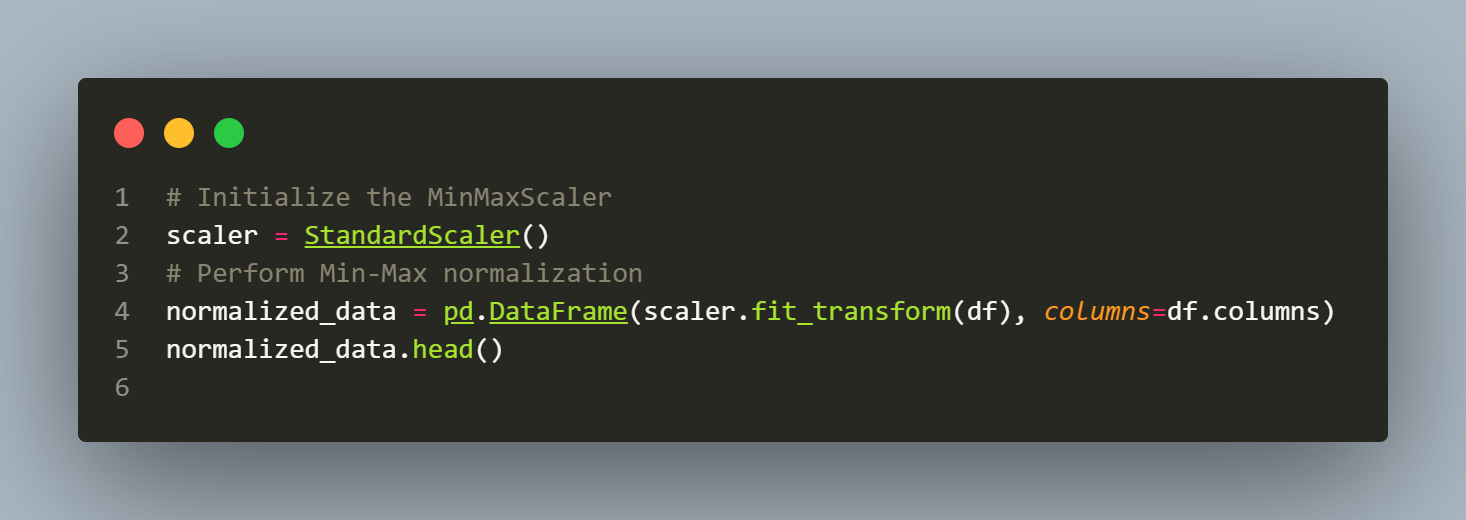
شکل 10: آمار توصیفی

آمار توصیفی نشان می‌دهد که ویژگی‌ها دارای مقیاس‌های مختلفی هستند:

* صادرات: در محدوده 0.109 تا 200.
* بهداشت: در محدوده 1.81 تا 17.9.
* واردات: در محدوده 0.0659 تا 174.
* درآمد: در محدوده 609 تا 125,000.
* تورم: در محدوده -4.21 تا 104.

این محدوده وسیع در مقادیر نشان می‌دهد که نرمال‌سازی لازم است تا همه ویژگی‌ها را به یک مقیاس مشابه بیاورد. این کار تضمین می‌کند که هیچ ویژگی خاصی به دلیل مقیاس خود بر تحلیل غالب نباشد.

ابتدا یک نمونه از کلاس StandardScaler را ایجاد می‌کنیم. این ابزار برای نرمال‌سازی داده‌ها به کار می‌رود. سپس نرمال‌سازی داده‌ها را با استفاده از StandardScaler انجام می‌دهیم. این فرآیند داده‌ها را به گونه‌ای تبدیل می‌کند که میانگین هر ویژگی برابر با 0 و انحراف معیار آن برابر با 1 باشد. در نهایت، چند سطر اول از داده‌های نرمال‌شده را نمایش می‌دهیم تا مطمئن شویم فرآیند نرمال‌سازی به درستی انجام شده است. شکل (11) نحوه انجام این کار را نمایش می‌دهد.

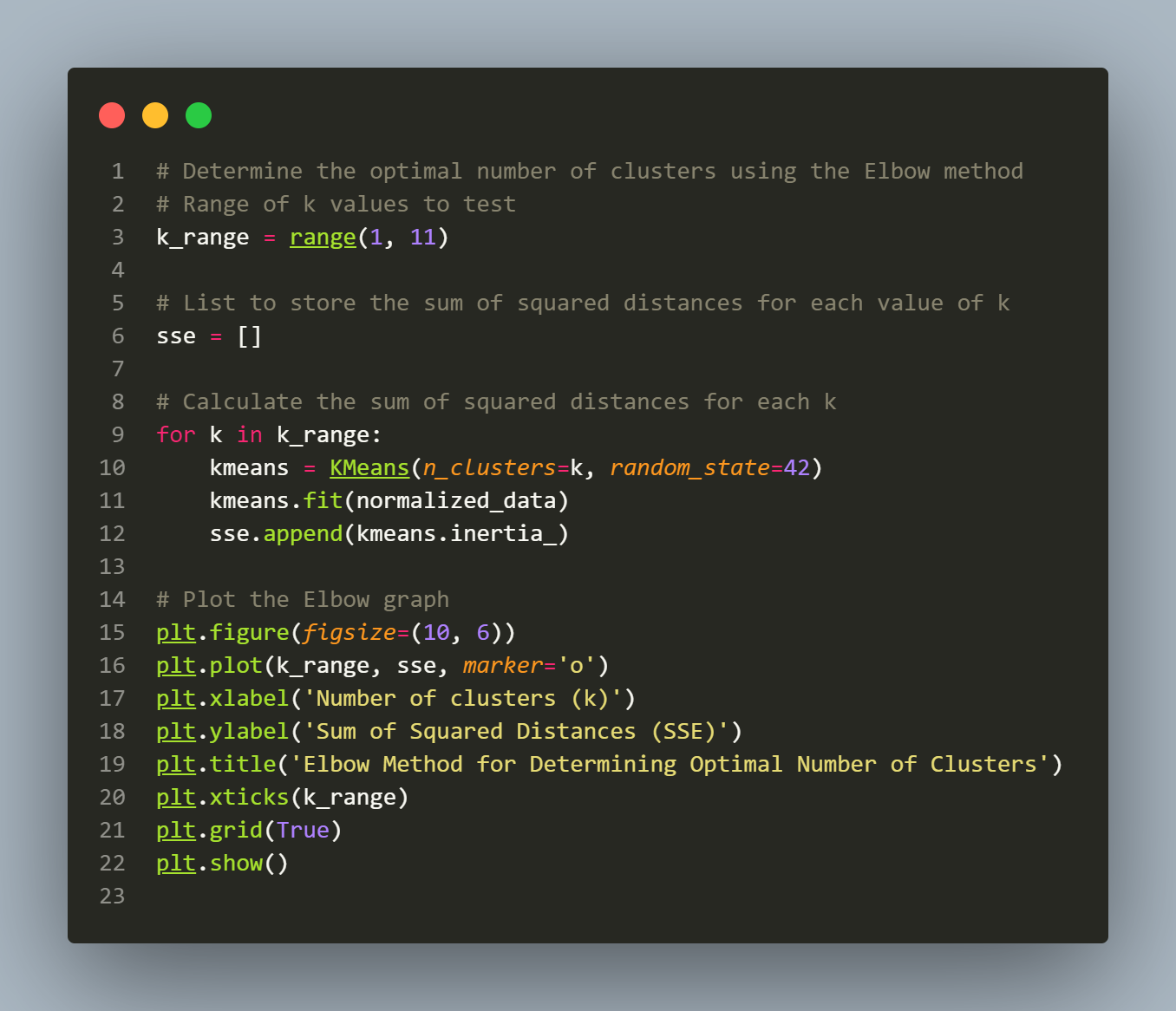


شکل 11: نرمال سازی داده‌ها

## 4-1-خوشه بندی

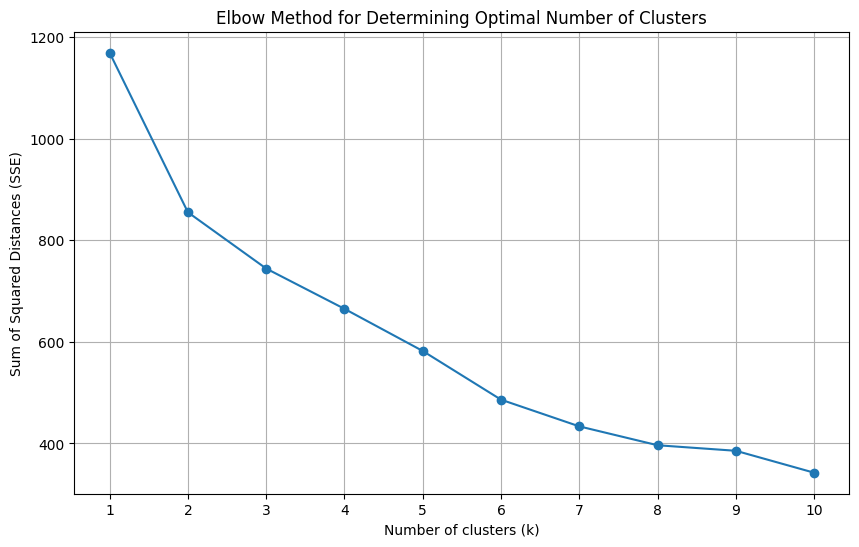
### 1-4-1-خوشه‌بندی با استفاده از K-means

ابتدا با روش آرنج به پیدا کردن K بهینه می‌پردازیم. ابتدا محدوده‌ای از مقادیر k که می‌خواهیم آزمایش کنیم را تعیین می‌کنیم. در اینجا، از 1 تا 10 استفاده می‌کنیم. لیستی ایجاد می‌کنیم تا مقادیر مجموع فواصل مربعی برای هر مقدار k را در آن ذخیره کنیم. برای هر مقدار k، مدل KMeans را با تعداد کلاسترهای k ایجاد و آموزش می‌دهیم و مجموع فواصل مربعی (inertia) را محاسبه و ذخیره می‌کنیم. ابتدا اندازه نمودار را تنظیم می‌کنیم. سپس نمودار Elbow را رسم می‌کنیم که در آن محور x تعداد کلاسترها و محور y مجموع فواصل مربعی است. همچنین، نمودار را به صورت نقطه‌چین نمایش می‌دهیم تا واضح‌تر باشد. شکل (12) نحوه انجام این کار را نمایش می‌دهد.



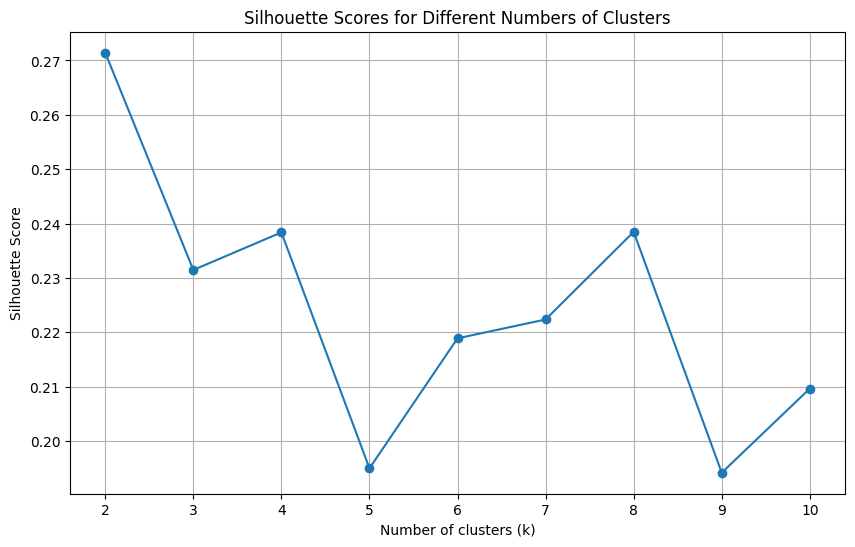
شکل 12: پیدا کردن کای بهینه با استفاده از روش آرنج

شکل (12) خروجی این کد را نمایش می‌دهد.



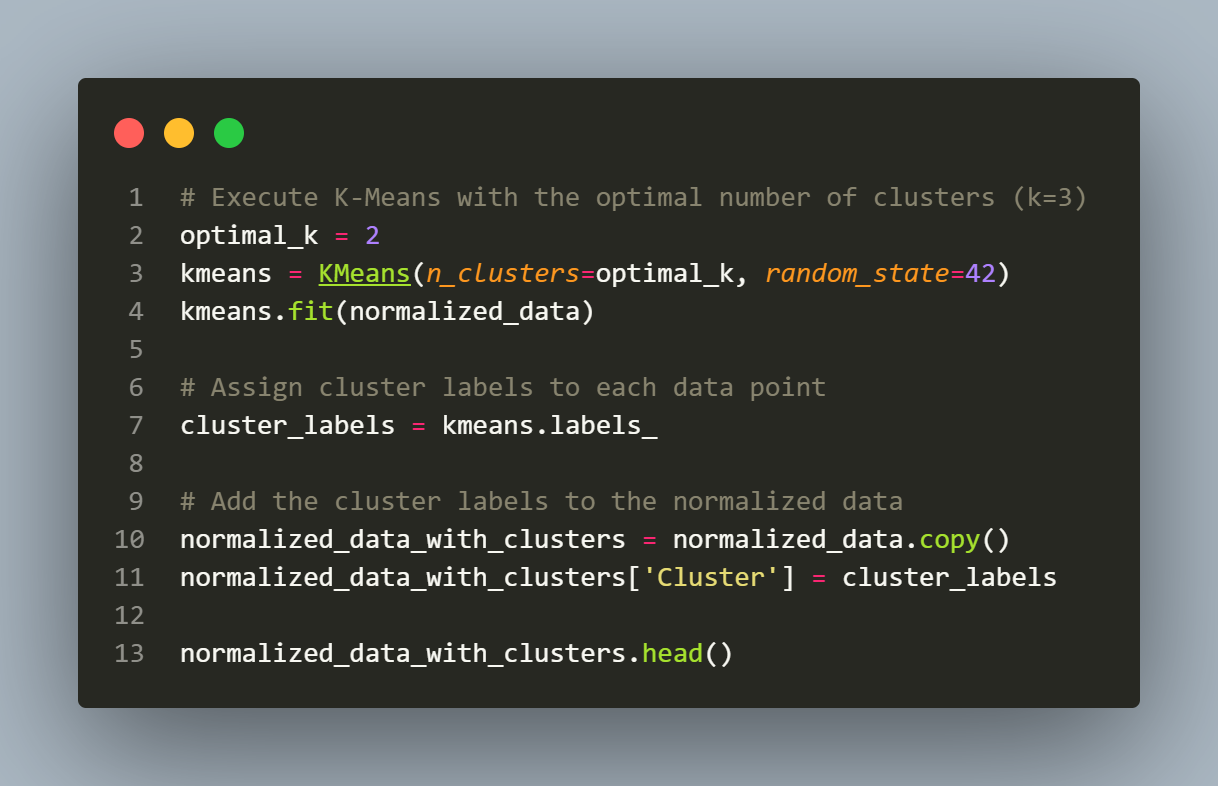
شکل 13: نمودار آرنج به دست آمده

سپس همان مراحل قبلی را تکرار می‌کنیم ولی به جای استفاده از مجموع فواصل مربعی، از امتیاز سیلوئت استفاده می‌کنیم. نمودار به دست آمده در شکل (14) نمایش داده شده است.



شکل 14: نمودار امتیاز سیلوئت

با توجه به امتیازهای به دست آمده به نظر می‌آید استفاده از دو خوشه مناسب باشد. مدل K-Means را با تعداد کلاسترهای بهینه ایجاد و آموزش می‌دهیم. برچسب‌های کلاستر که توسط مدل K-Means تخصیص داده شده‌اند را به هر نقطه داده اختصاص می‌دهیم. ابتدا یک نسخه کپی از داده‌های نرمال‌شده ایجاد می‌کنیم و سپس برچسب‌های کلاستر را به عنوان یک ستون جدید به این داده‌ها اضافه می‌کنیم. در نهایت، چند سطر اول از داده‌های نرمال‌شده که شامل برچسب‌های کلاستر هستند را نمایش می‌دهیم تا مطمئن شویم که برچسب‌ها به درستی اضافه شده‌اند. شکل (15) نحوه انجام این کار را نمایش می‌دهد.



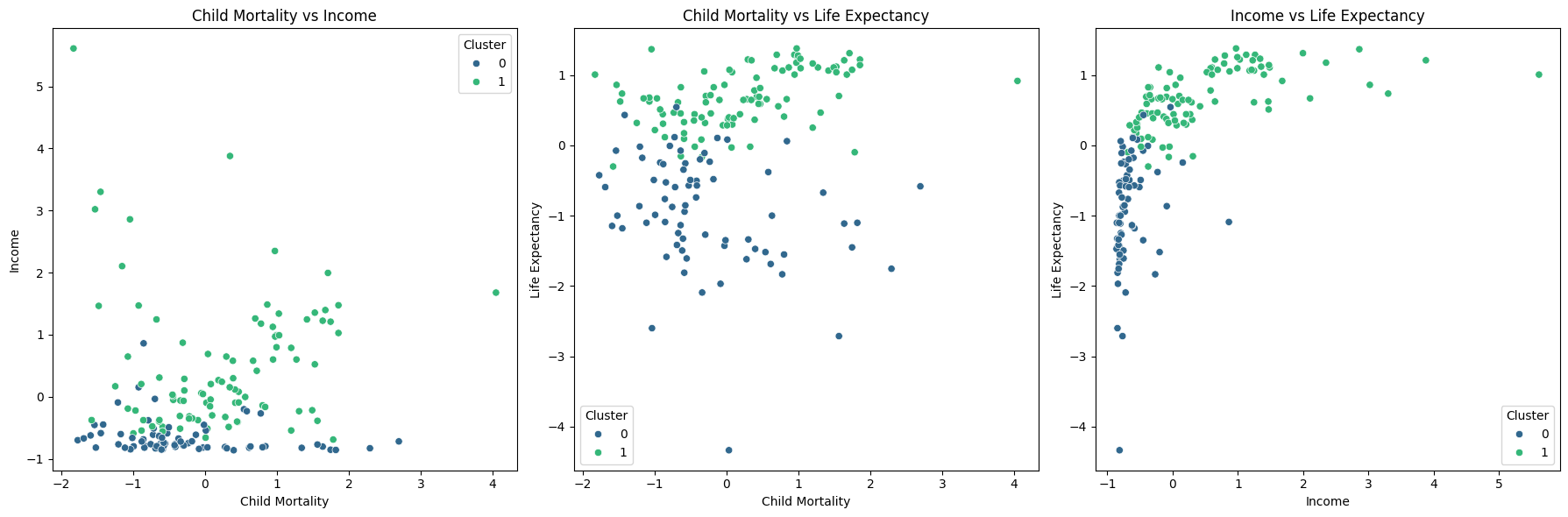
شکل 15: خوشه‌بندی با تعداد خوشه بهینه

همچنین خروجی این کد در شکل (16) آمده.



شکل 16: دیتاست برچسب دار شده

ابتدا سه ویژگی 'health', 'income' و 'life\_expec' را برای نمودارهای پراکندگی انتخاب می‌کنیم. اندازه کلی نمودارها را تنظیم می‌کنیم تا هر سه نمودار به خوبی در کنار هم قرار بگیرند. یک subplot جدید برای نمودار 'health' در مقابل 'income' ایجاد می‌کنیم. داده‌ها را با استفاده از sns.scatterplot رسم می‌کنیم و از برچسب‌های کلاسترها برای تعیین رنگ نقاط استفاده می‌کنیم. مشابه این کار را برای دو نمودار دیگر انجام می‌دهیم. با استفاده از plt.tight\_layout() اطمینان حاصل می‌کنیم که نمودارها به خوبی در کنار هم قرار گرفته‌اند و هیچ تداخل بصری ندارند. سپس نمودارها را نمایش می‌دهیم. شکل (17) این سه نمودار را نمایش می‌دهد.



شکل 17: نمودار پرداکندی برای سه ستون

* نمودار (Child Mortality vs Income):

محور افقی نشان‌دهنده مرگ و میر کودکان و محور عمودی نشان‌دهنده درآمد است.

در این نمودار، دو کلاستر با رنگ‌های مختلف نشان داده شده‌اند.

نقاط کلاستر 0 (رنگ آبی) در بخش‌های پایین و متمرکز به سمت چپ نمودار بیشتر مشاهده می‌شوند، که نشان‌دهنده درآمد کمتر و مرگ و میر کودکان بیشتر است.

نقاط کلاستر 1 (رنگ سبز) پراکندگی بیشتری دارند و به نظر می‌رسد که بیشتر در بخش‌های بالاتر نمودار قرار گرفته‌اند، که نشان‌دهنده درآمد بالاتر و مرگ و میر کودکان کمتر است.

* نمودار (Child Mortality vs Life Expectancy):

محور افقی نشان‌دهنده مرگ و میر کودکان و محور عمودی نشان‌دهنده امید به زندگی است.

نقاط کلاستر 0 (رنگ آبی) به طور کلی در پایین نمودار متمرکز شده‌اند، که نشان‌دهنده امید به زندگی کمتر و مرگ و میر کودکان بیشتر است.

نقاط کلاستر 1 (رنگ سبز) بیشتر در قسمت‌های بالای نمودار قرار دارند، که نشان‌دهنده امید به زندگی بیشتر و مرگ و میر کودکان کمتر است.

* نمودار سمت راست (Income vs Life Expectancy):

محور افقی نشان‌دهنده درآمد و محور عمودی نشان‌دهنده امید به زندگی است.

نقاط کلاستر 0 (رنگ آبی) بیشتر در بخش‌های پایین و متمرکز به سمت چپ نمودار قرار دارند، که نشان‌دهنده درآمد کمتر و امید به زندگی کمتر است.

نقاط کلاستر 1 (رنگ سبز) پراکندگی بیشتری دارند و به نظر می‌رسد که بیشتر در بخش‌های بالاتر نمودار قرار گرفته‌اند، که نشان‌دهنده درآمد بالاتر و امید به زندگی بیشتر است.

نتیجه‌گیری:

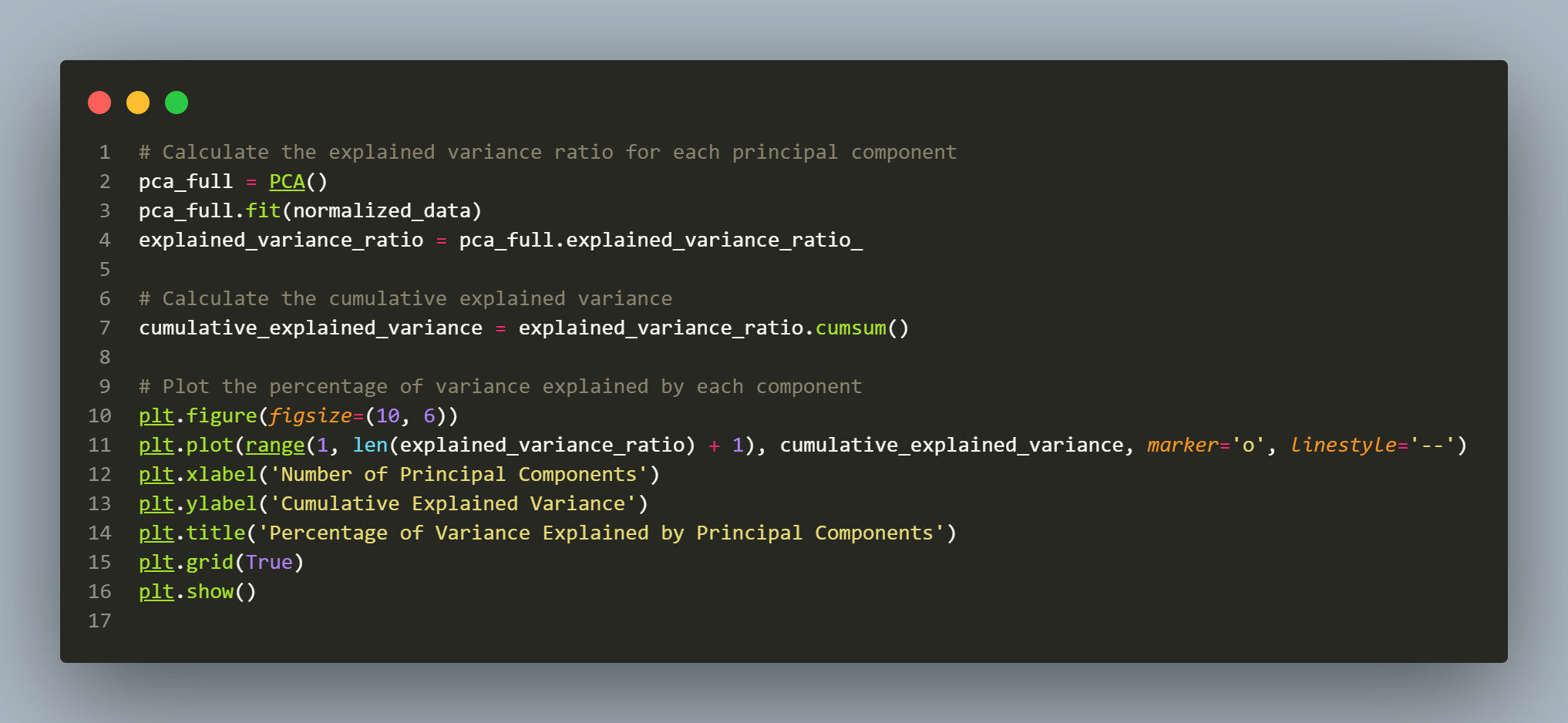
به نظر می‌رسد که دو کلاستر مختلف از داده‌ها وجود دارند که تفاوت‌های چشمگیری در متغیرهای مورد بررسی دارند.

کلاستر 0 (رنگ آبی) نمایانگر کشورها یا مناطقی با درآمد کمتر، مرگ و میر کودکان بیشتر و امید به زندگی کمتر است.

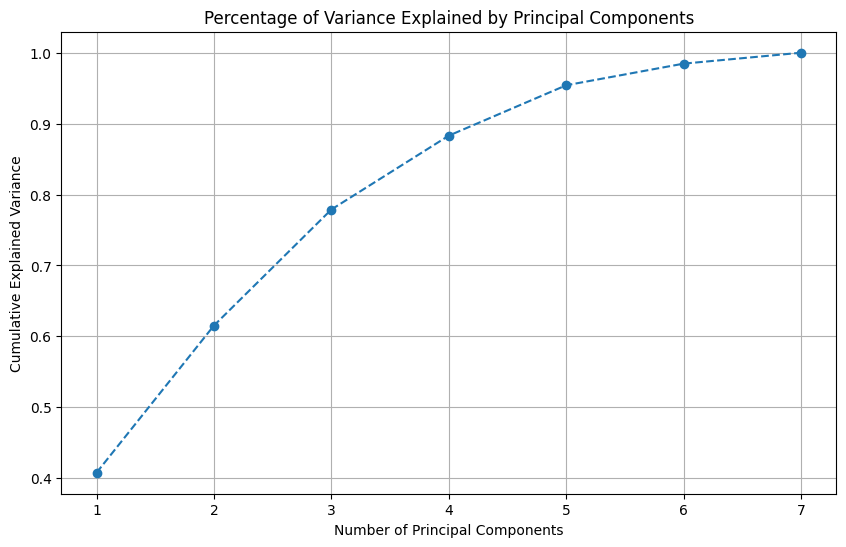
کلاستر 1 (رنگ سبز) نمایانگر کشورها یا مناطقی با درآمد بالاتر، مرگ و میر کودکان کمتر و امید به زندگی بیشتر است.

## 5-1-استفاده از PCA

ابتدا یک نمونه از کلاس PCA بدون محدودیت تعداد مؤلفه‌ها ایجاد می‌کنیم و داده‌های نرمال‌شده را با آن تناسب می‌دهیم تا نسبت واریانس توضیح داده شده توسط هر مؤلفه اصلی محاسبه شود. نسبت واریانس توضیح داده شده توسط هر مؤلفه اصلی را محاسبه می‌کنیم. واریانس تجمعی توضیح داده شده توسط مؤلفه‌ها را محاسبه می‌کنیم. سپس نمودار نسبت واریانس توضیح داده شده تجمعی را رسم می‌کنیم. در این نمودار محور x تعداد مؤلفه‌های اصلی و محور y نسبت واریانس توضیح داده شده تجمعی است. شکل (18) کد و شکل (19) خروجی کد را نمایش می‌دهد.



شکل 18: کد مربوط به رسم نمودار PCA



شکل 19: نمودار میزان واریانس قابل توضیح توسط هر PC