گزارش فاز اول پروژه مبانی داده کاوی

هدف : تحلیل آماری اولیه و کاهش بعد داده‌ها

مبینا صولتی ۴۰۰۱۲۶۲۰۶۸

گروه ۴ دیتاست : adult

۱ .تحلیل آماری اولیه

محاسبه ماتریس کوواریانس:

برای بررسی همبستگی بین ویژگی‌ها، ماتریس کوواریانس محاسبه شد:

ماتریس کوواریانس ابزاری است برای سنجش میزان تغییر مشترک بین ویژگی‌ها. هرچه مقدار کوواریانس بین دو ویژگی بزرگ‌تر (مثبت یا منفی) باشد، آن‌ها همبستگی بیشتری دارند.

```
raw matrix:
                      fnlwgt  capital-loss  hours-per-week
fnlwgt          4.365972e+11 -2.747481e+06   -17653.467509
capital-loss   -2.747481e+06  1.789419e+05      310.470076
hours-per-week -1.765347e+04  3.104701e+02    17520.034606

matrix shape: (3, 3)
```

توضیح ماتریس کوواریانس :

- **ماتریس کوواریانس چه اطلاعاتی می‌دهد؟** :

مقادیر در ماتریس نشان‌دهنده ضرایب همبستگی بین این ویژگی‌ها هستند:

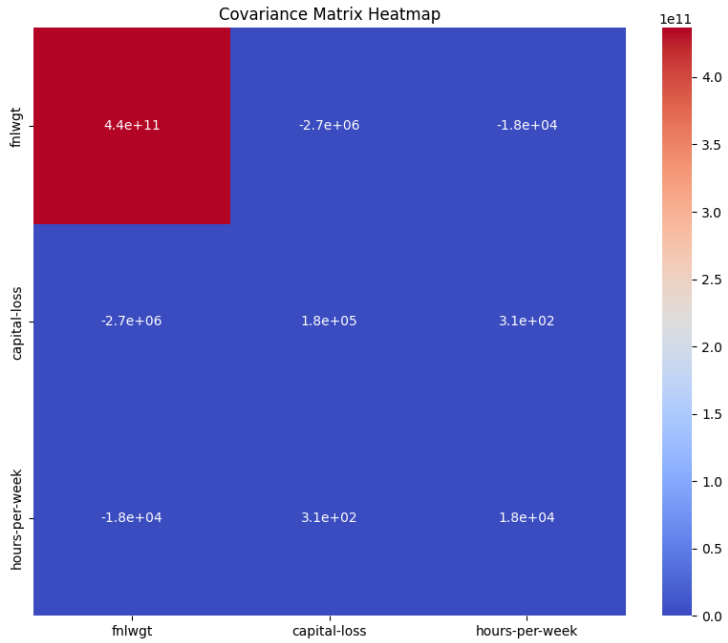
- مقادیر نزدیک به ۱ یا -۱ نشان‌دهنده همبستگی قوی (مثبت یا منفی) هستند.
- مقدار ۰ نشان‌دهنده عدم همبستگی است.

بر اساس داده‌ها:

- همبستگی fnlwgtبا خودش ۱۱e+۴.۳۶۵۹۷۲ است (که به دلیل مقیاس بزرگ ممکن است خطا یا مقیاس‌بندی خاص باشد).
- همبستگی fnlwgt با capital-loss و با hours-per-week منفی است، که نشان‌دهنده رابطه معکوس ضعیف تا متوسط است.
- همبستگی capital-lossبا خودش ۰۶e+۱.۷۸۴۹۱ و با hours-per-week۳۱۰.۴۷۰۰۷۶ است، که نشان‌دهنده رابطه مثبت ضعیف است.
- همبستگی hours-per-weekبا خودش ۱۷۵۲۹.۰۳۴۰۶ است.

این ماتریس می‌تواند برای تحلیل اینکه چگونه تغییرات در یک ویژگی (مثلاً ساعت کاری) با تغییرات در ویژگی‌های دیگر (مثلاً سرمایه از دست رفته) مرتبط است، استفاده شود. با این حال، مقادیر غیرمعمولاً بزرگ ممکن است نیاز به بررسی بیشتر یا نرمال‌سازی داده‌ها داشته باشد.

تصویری Heatmap:

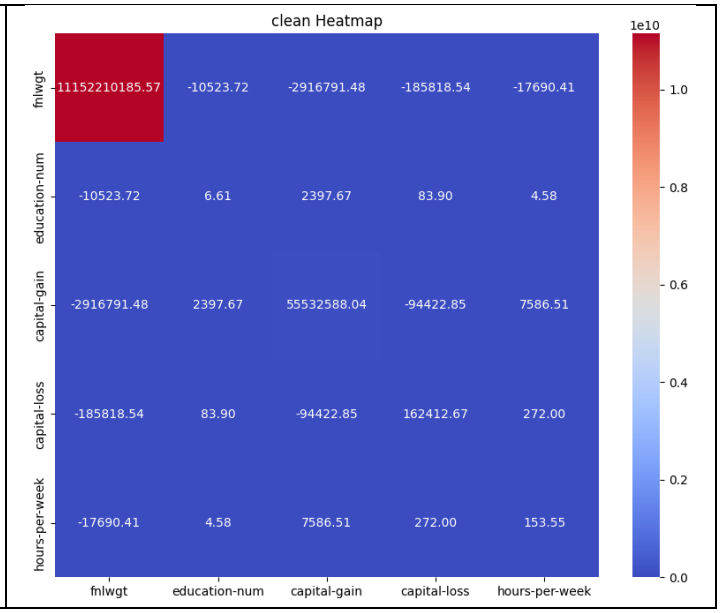


برای دیتاست تمیز:

```
clean matrix:
                    fnlwgt  education-num  capital-gain   capital-loss  \
fnlwgt          1.115221e+10  -10523.722187 -2.916791e+06 -185818.540554
education-num  -1.052372e+04       6.609901  2.397673e+03      83.896136
capital-gain   -2.916791e+06    2397.673247  5.553259e+07  -94422.847037
capital-loss   -1.858185e+05      83.896136 -9.442285e+04  162412.669033
hours-per-week -1.769041e+04       4.577651  7.586508e+03     271.998838

                hours-per-week
fnlwgt           -17690.405781
education-num        4.577651
capital-gain      7586.507972
capital-loss       271.998838
hours-per-week     153.547885

matrix shape: (5, 5)
```



ابعاد ماتریس

ابعاد این ماتریس (۵، ۵) است، همان‌طور که در انتهای تصویر مشخص شده .((matrix shape: (5, 5) این یعنی ماتریس دارای ۵ ردیف و ۵ ستون است، که هر ردیف و ستون به ترتیب به یکی از ویژگی‌هاfnlwgt، education-num، capital-gain، capital-loss، hours-per-week مربوط می‌شود.

اطلاعات ماتریس درباره ارتباط بین ویژگی‌ها

این ماتریس یک **ماتریس کوواریانس** است که ارتباط بین ویژگی‌های داده‌شده را نشان می‌دهد. در اینجا، هر مقدار در ماتریس نشان‌دهنده میزان ارتباط (کوواریانس) بین دو ویژگی است:

۱. **قطر اصلی (دیاگونال) :**مقادیر روی قطر اصلی) مانند ۱.۱۱۵۲۲۱e+۱۰ برای fnlwgt یا ۵.۵۳۳۲۵۹e+۰۷ برای capital-gain)واریانس هر ویژگی را نشان می‌دهند. واریانس بالا) مانند ۱.۱۱۵۲۲۱e+۱۰ برای (fnlwgt نشان می‌دهد که داده‌های این ویژگی پراکندگی زیادی دارند.

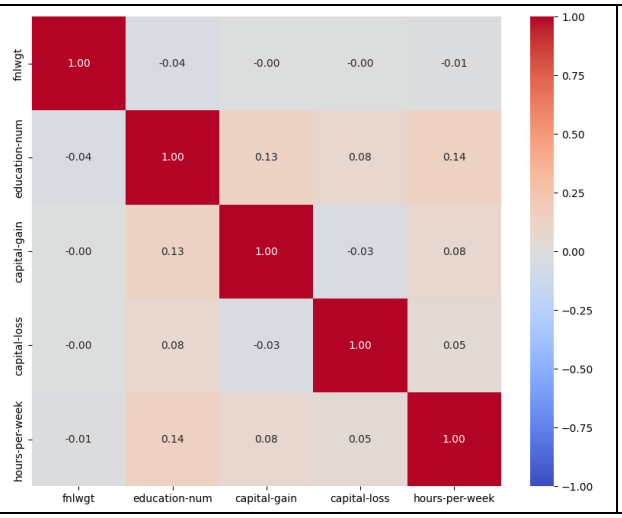
۲. **مقادیر خارج از قطر :**این مقادیر کوواریانس بین دو ویژگی را نشان می‌دهند. برای مثال :

- کوواریانس بین capital-gain و capital-loss برابر با -۹.۴۴۲۲۸۵e+۰۴ است (منفی). این نشان می‌دهد که این دو ویژگی رابطه معکوس دارند؛ یعنی وقتی capital-gain افزایش می‌یابد، capital-lossتمایل به کاهش دارد و بالعکس.
- کوواریانس بین education-num و hours-per-week برابر با ۴.۵۷۷۶۵۱ است (مثبت). این نشان می‌دهد که این دو ویژگی رابطه مستقیم دارند؛ یعنی افرادی با سطح تحصیلات بالاتر (education-num بیشتر (به طور متوسط ساعات کاری بیشتری (hours-per-week) دارند.

۳. **قدرت و جهت ارتباط :**

- مقادیر مثبت) مانند ۲.۳۹۷۶۷۳e+۰۳ بین education-num و (capital-gain نشان‌دهنده رابطه مستقیم هستند.
- مقادیر منفی) مانند -۱.۷۶۹۰۴۱e+۰۴ بین fnlwgt و (hours-per-week نشان‌دهنده رابطه معکوس هستند.
- هرچه قدر مطلق یک مقدار بیشتر باشد، ارتباط بین دو ویژگی قوی‌تر است. برای مثال، کوواریانس -۹.۴۴۲۲۸۵e+۰۴ بین capital-gain و capital-loss نشان‌دهنده یک رابطه معکوس نسبتاً قوی است.

تحلیل همبستگی




```
corr matrix for clean:
                fnlwgt  education-num  capital-gain  capital-loss  \
fnlwgt        1.000000      -0.038761     -0.003706     -0.004366
education-num -0.038761      1.000000      0.125146      0.080972
capital-gain  -0.003706      0.125146      1.000000     -0.031441
capital-loss  -0.004366      0.080972     -0.031441      1.000000
hours-per-week -0.013519     0.143689      0.082157      0.054467

                hours-per-week
fnlwgt               -0.013519
education-num         0.143689
capital-gain          0.082157
capital-loss          0.054467
hours-per-week        1.000000
```

این تصویر یک ماتریس همبستگی (Correlation Matrix)را نشان می‌دهد که میزان همبستگی بین ویژگی‌هایfnlwgt ، education-num، capital-gain، capital-lossو hours-per-week را بررسی می‌کند. همبستگی مقادیر بین -۱ (همبستگی معکوس کامل) و ۱ (همبستگی مستقیم کامل) دارد، و مقدار ۰ نشان‌دهنده عدم همبستگی است.

۱ .بررسی همبستگی‌های بالا یا شباهت زیاد

- همبستگی‌های قابل توجه :
  - education-numو hours-per-week: همبستگی ۰.۱۴۳۶۸۹ (مثبت). این نشان می‌دهد که با افزایش سطح تحصیلات(education-num) ، ساعات کاری در هفته (hours-per-week) نیز به طور متوسط افزایش می‌یابد. این همبستگی مثبت است اما مقدار آن نسبتاً پایین است (کمتر از ۰.۲)، پس نمی‌توان آن را "بسیار بالا" در نظر گرفت.
  - education-numو capital-gain: همبستگی ۰.۱۲۵۱۴۶ (مثبت). این نشان می‌دهد که افراد با تحصیلات بالاتر به طور متوسط سود سرمایه بیشتری دارند، اما باز هم این مقدار نسبتاً پایین است.
  - capital-gainو capital-loss: همبستگی -۰.۰۳۱۴۴۱ (منفی). این نشان‌دهنده یک رابطه معکوس ضعیف است؛ یعنی وقتی سود سرمایه افزایش می‌یابد، زیان سرمایه کمی کاهش می‌یابد.
  - capital-lossو hours-per-week: همبستگی ۰.۰۵۴۴۶۷ (مثبت). این مقدار بسیار پایین است و نشان‌دهنده یک رابطه مستقیم ضعیف است.

۲ .آیا ویژگی‌هایی با همبستگی بسیار بالا یا شباهت زیاد وجود دارند؟

- همبستگی بسیار بالا معمولاً به مقادیر نزدیک به ۱ یا -۱ (مثلاً بالاتر از ۰.۷ یا کمتر از -۰.۷) گفته می‌شود. در این ماتریس، هیچ جفت ویژگی‌ای چنین همبستگی بالایی ندارد. بالاترین همبستگی بین education-num وhours-per-week۰.۱۴۳۶۸۹است که همچنان ضعیف محسوب می‌شود.
- بنابراین، هیچ دو ویژگی‌ای شباهت بسیار زیاد یا همبستگی قوی ندارند که بتوان آن‌ها را تکراری در نظر گرفت.

۳ .آیا می‌توان ویژگی‌ها را به عنوان تکراری یا غیرضروری حذف کرد؟

- ویژگی‌های تکراری :از آنجایی که هیچ همبستگی بسیار بالایی (مثلاً بالای ۰.۷) بین ویژگی‌ها وجود ندارد، نمی‌توان هیچ ویژگی‌ای را به عنوان "تکراری" حذف کرد. ویژگی‌های تکراری معمولاً آن‌هایی هستند که اطلاعات مشابهی ارائه می‌دهند (همبستگی نزدیک به ۱ یا -۱)، اما در اینجا چنین وضعیتی مشاهده نمی‌شود.
- ویژگی‌های غیرضروری :ویژگی fnlwgt با تمام ویژگی‌های دیگر همبستگی بسیار پایینی (نزدیک به صفر) دارد. این ممکن است نشان‌دهنده این باشد که fnlwgt اطلاعات زیادی درباره تغییرات سایر ویژگی‌ها ارائه نمی‌دهد و شاید در برخی مدل‌ها (مثلاً مدل‌های پیش‌بینی) نقش مهمی نداشته باشد. با این حال، تصمیم به حذف آن بستگی به هدف تحلیل دارد :
  - اگر هدف پیش‌بینی یا مدل‌سازی است، می‌توان با آزمایش مثلاً حذف fnlwgt و بررسی عملکرد مدل تصمیم گرفت که آیا این ویژگی غیرضروری است یا خیر.
  - اگر هدف تحلیل اکتشافی داده‌هاست، بهتر است fnlwgt را نگه داشت، زیرا ممکن است در تحلیل‌های دیگر (مثلاً با متغیرهای دیگر یا در مدل‌های غیرخطی) مفید باشد.

۴ .تحلیل کلی

- روابط ضعیف :بیشتر همبستگی‌ها در این ماتریس ضعیف هستند (کمتر از |۰.۲|). این نشان می‌دهد که ویژگی‌ها به طور کلی مستقل از یکدیگر عمل می‌کنند و اطلاعات متفاوتی ارائه می‌دهند.
- روابط معنی‌دار :تنها روابطی که کمی قابل توجه هستند (هرچند ضعیف) بین education-num و-hours-per week) ۰.۱۴۳۶۸۹ (و education-num و) capital-gain ۰.۱۲۵۱۴۶ (دیده می‌شود. این روابط منطقی به نظر می‌رسند: تحصیلات بالاتر معمولاً با ساعات کاری بیشتر و سود سرمایه بیشتر همراه است.
- capital-gainو capital-loss: همبستگی منفی ضعیف (-۰.۰۳۱۴۴۱) بین این دو ویژگی نشان می‌دهد که این دو متغیر تا حدی معکوس عمل می‌کنند، اما این رابطه آنقدر قوی نیست که یکی را بتوان جایگزین دیگری کرد.

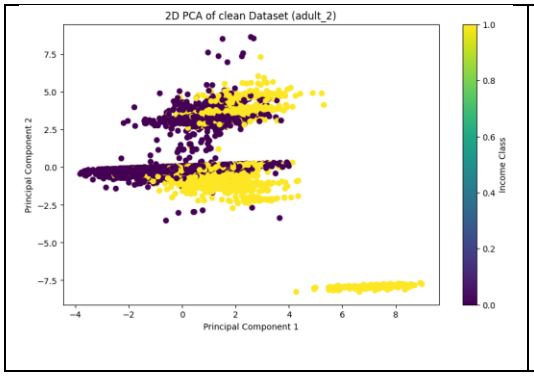
نتیجه‌گیری

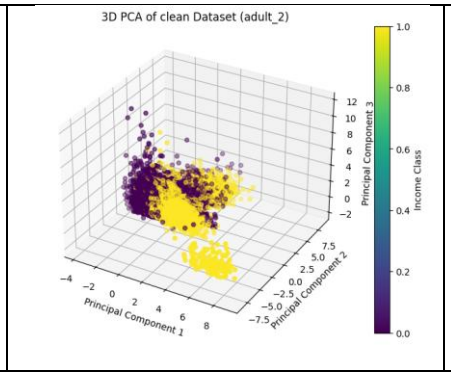
- هیچ ویژگی‌ای همبستگی بسیار بالا یا شباهت زیاد با دیگری ندارد، بنابراین نمی‌توان ویژگی‌ای را به عنوان "تکراری" حذف کرد.
- fnlwgtبه دلیل همبستگی بسیار پایین با سایر ویژگی‌ها ممکن است در برخی مدل‌ها غیرضروری به نظر برسد، اما حذف آن باید با آزمایش و تحلیل بیشتر (مثلاً بررسی اهمیت ویژگی در یک مدل یادگیری ماشین) تأیید شود.
- سایر ویژگی‌هاeducation-num) ، capital-gain، capital-loss، hours-per-week)اطلاعات متفاوتی ارائه می‌دهند و بهتر است در تحلیل نگه داشته شوند، مگر اینکه تحلیل بیشتری خلاف آن را نشان دهد.

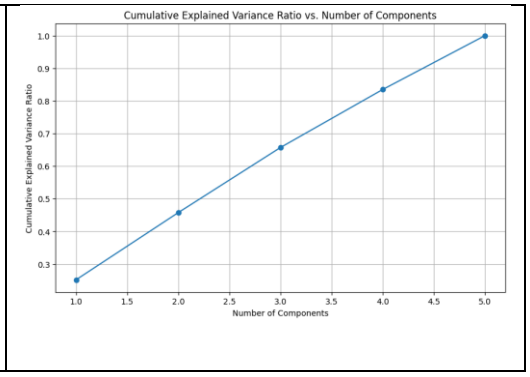
۲ .کاهش ابعاد داده ها

در این بخش، سه روش معروف کاهش بعد اجرا شد:

روش PCA تحلیل مؤلفه‌های اصلی







- مراحل انجام:
  - محاسبه کوواریانس: ابتدا ماتریس کوواریانس داده‌ها محاسبه می‌شود تا رابطه بین ویژگی‌ها مشخص شود.
  - محاسبه مقادیر ویژه و بردارهای ویژه: مقادیر ویژه (eigenvalues) و بردارهای ویژه (eigenvectors) ماتریس کوواریانس به دست می‌آید که نشان‌دهنده جهت و میزان واریانس هستند.

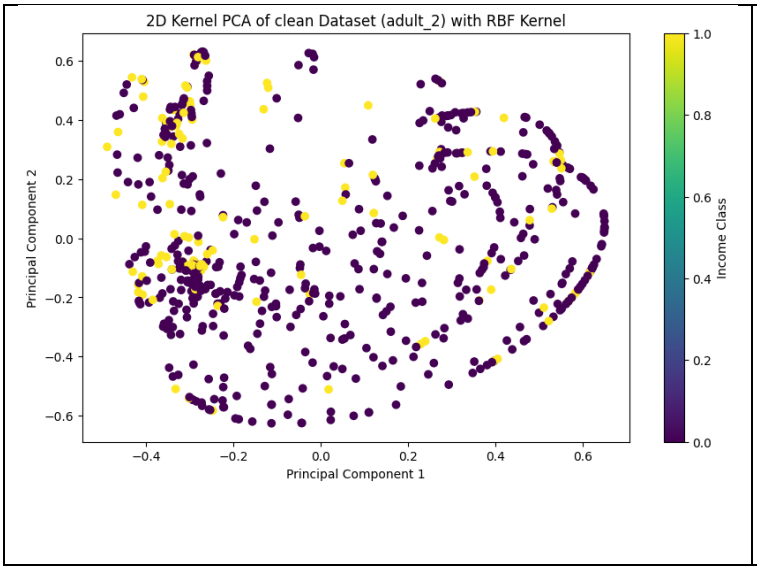
- انتخاب مؤلفه‌ها: مؤلفه‌هایی که بیشترین واریانس را توضیح می‌دهند انتخاب می‌شوند. در اینجا، نمودار "Cumulative Explained Variance Ratio" نشان می‌دهد که با افزایش تعداد مؤلفه‌ها، چه مقدار از واریانس کل حفظ می‌شود.
- تبدیل داده‌ها: داده‌ها به فضای جدید با مؤلفه‌های انتخاب‌شده (در اینجا D۲) پروژه می‌شوند.

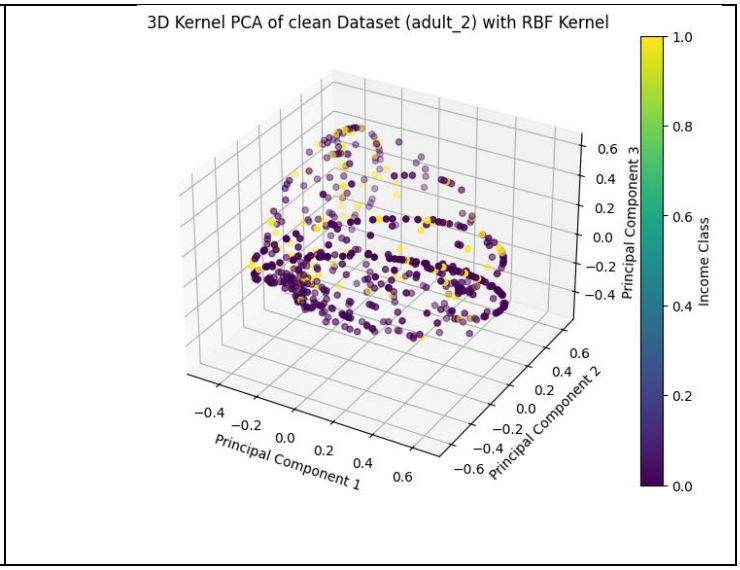
- دلیل انتخاب ابعاد (D۲):

در نمودار "Cumulative Explained Variance Ratio"، با ۲ مؤلفه اصلی حدود ۰.۶ تا ۰.۷ واریانس داده‌ها حفظ می‌شود (بسته به آستانه انتخابی). این یعنی با استفاده از ۲ مؤلفه، بخش قابل توجهی از اطلاعات حفظ می‌شود، هرچند همه واریانس (۱۰۰%) پوشش داده نمی‌شود.

در نمودار D PCA۲، دو خوشه (بنفش و زرد) به خوبی از هم جدا شده‌اند که نشان می‌دهد ۲ بعد برای تفکیک کلاس‌های درآمد (income class) کافی است. این مصورسازی نشان می‌دهد که PCA داده‌ها را به صورت خطی کاهش داده و ساختار کلی را حفظ کرده است.

روش Kernel PCA





- مراحل انجام:

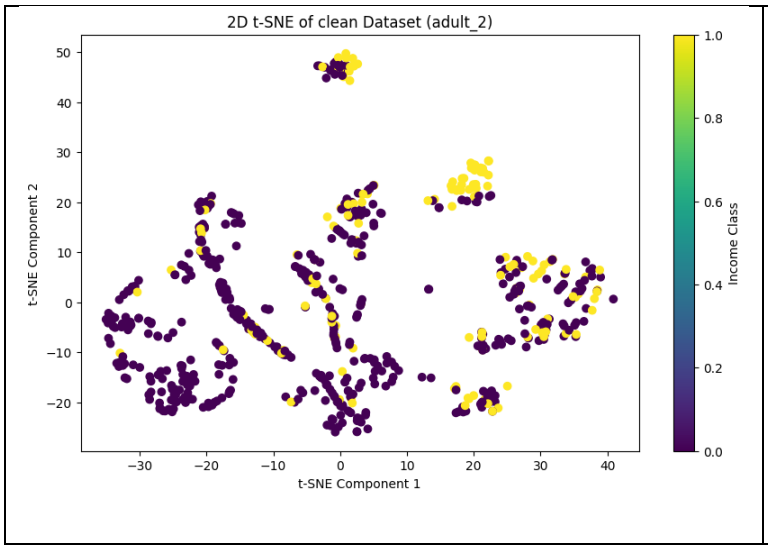
- نقشه‌برداری به فضای بالاتر: با استفاده از کرنل RBF (Radial Basis Function)، داده‌ها به یک فضای غیرخطی با ابعاد بالاتر منتقل می‌شوند.
- انجام PCA در فضای کرنلی: در این فضای جدید، PCA اعمال می‌شود تا مؤلفه‌های اصلی استخراج شوند.
- کاهش به D۲: داده‌ها به ۲ بعد کاهش می‌یابند و برای مصورسازی استفاده می‌شوند.

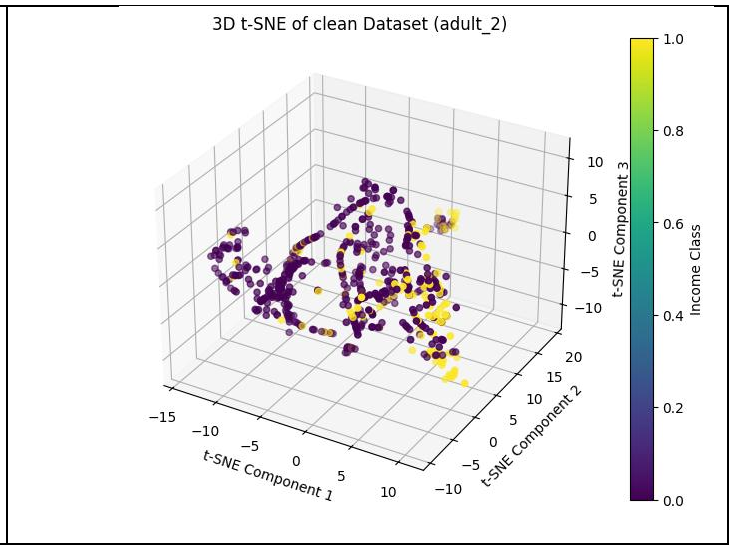
- دلیل انتخاب ابعاد (D۲):

مانند PCA، انتخاب D۲ بر اساس نمودار واریانس تجمعی است. با ۲ مؤلفه، بخشی از واریانس (حدود ۰.۶ تا ۰.۷) حفظ می‌شود، اما چون Kernel PCA غیرخطی است، می‌تواند الگوهای پیچیده‌تر را نسبت به PCA ساده بهتر نشان دهد.

در نمودار D Kernel PCA۲، پراکندگی داده‌ها کمی متفاوت است و خوشه‌ها (بنفش و زرد) به صورت غیرخطی جدا شده‌اند. این نشان می‌دهد که ۲ بعد برای حفظ ساختار غیرخطی داده‌ها کافی است، هرچند جداسازی کامل نیست.

روشt-SNE





- مراحل انجام:
  - محاسبه شباهت‌ها: فاصله‌های محلی بین نقاط داده در فضای اصلی محاسبه می‌شود و به صورت احتمالاتی مدل می‌شود.
  - بهینه‌سازی: با استفاده از تابع هزینه (cost function)، داده‌ها در فضای D۲ به گونه‌ای جابه‌جا می‌شوند که شباهت‌های محلی حفظ شود.
  - مصورسازی: نتیجه به صورت D۲ یا D۳ نمایش داده می‌شود.
- دلیل انتخاب ابعاد (D۲):

t-SNE به طور خاص برای مصورسازی طراحی شده و معمولاً D۲ یا D۳ انتخاب می‌شود، چون هدف آن نمایش ساختار محلی داده‌هاست، نه حفظ واریانس کل. در نمودار D t-SNE۲، خوشه‌های کوچک‌تر و جزئیات بیشتری دیده می‌شود که نشان‌دهنده حفظ روابط محلی است.

در اینجا، ۲ بعد برای تفکیک خوشه‌ها (بنفش و زرد) کافی به نظر می‌رسد، هرچند برخی نقاط پراکنده هستند که نشان می‌دهد t-SNE ممکن است بهینه‌سازی جهانی را تضمین نکند.

آیا واریانس می‌تواند معیار مناسبی برای سنجش میزان اطلاعات حفظ شده در فرایند کاهش باشد؟

- **پاسخ: بله، اما با محدودیت‌ها.**
- **دلایل با استناد به مفاهیم آماری :**

۱. **مفهوم واریانس :واریانس نشان‌دهنده پراکندگی داده‌هاست. درPCA ، حفظ واریانس کل به این معناست که اطلاعات اصلی داده‌ها (تفاوت‌ها بین نقاط) تا حد ممکن نگه داشته شود. مثلاً در نمودار واریانس تجمعی، با ۲ مؤلفه حدود ۶۰-۷۰% واریانس حفظ شده، که نشان‌دهنده حفظ بخش مهمی از اطلاعات خطی است.**

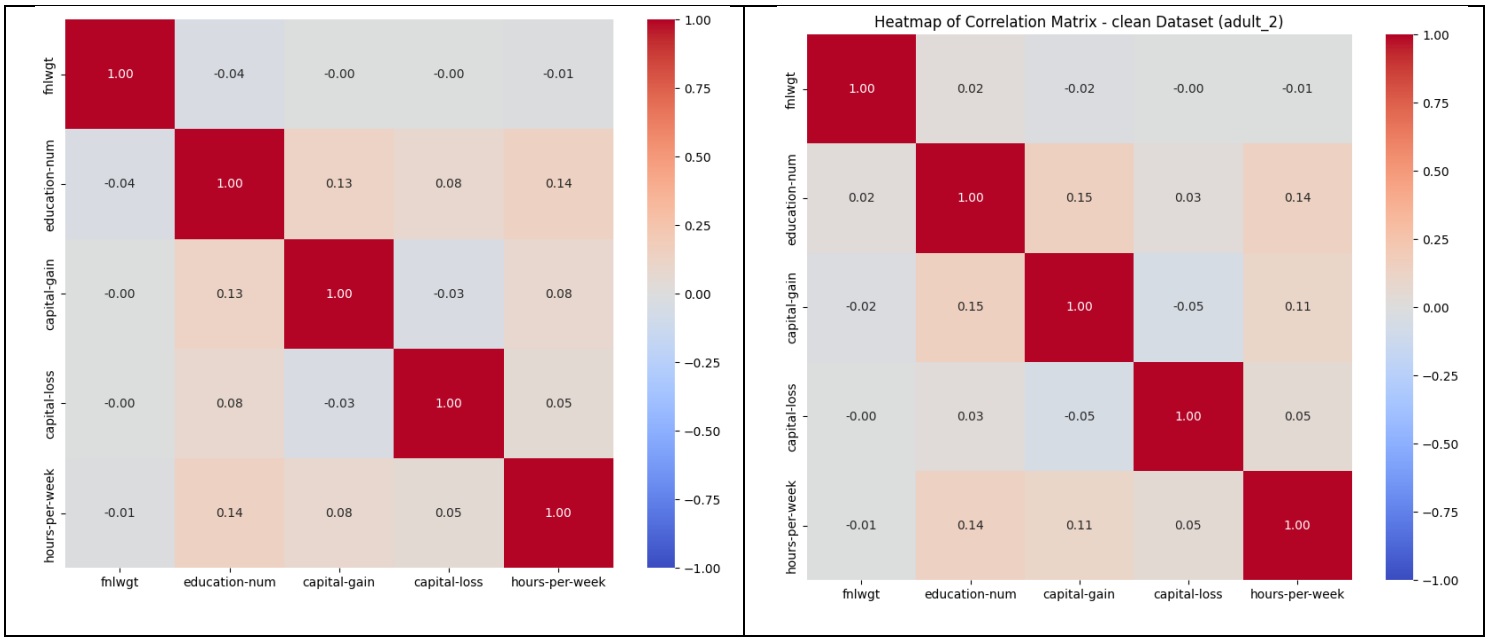
۲. **محدودیت برای ساختارهای غیرخطی :واریانس فقط اطلاعات خطی را اندازه‌گیری می‌کند. در داده‌های پیچیده (مثل خوشه‌های غیرخطی)، مثل آنچه در Kernel PCA و t-SNE دیده می‌شود، واریانس ممکن است کل اطلاعات (مثل روابط محلی) را نشان ندهد. مثلاً در ۲D t-SNE، خوشه‌ها به خوبی جدا شده‌اند، اما واریانس کل ممکن است پایین باشد.**

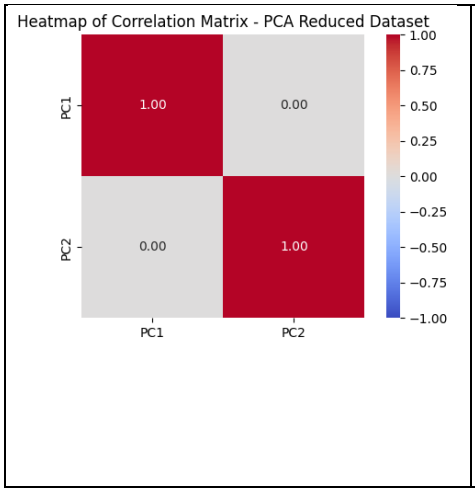
۳. **مثال از خروجی :درPCA ، ۲ Dحدود ۷۰% واریانس را حفظ می‌کند و خوشه‌ها قابل تشخیص‌اند. اما در t-SNE، جداسازی خوشه‌ها بهتر است، هرچند واریانس به عنوان معیار اصلی استفاده نمی‌شود، چون t-SNEروی حفظ شباهت‌های محلی تمرکز دارد، نه واریانس کل.**

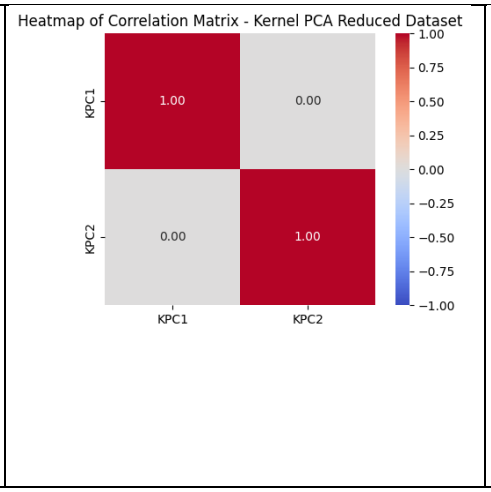
- **نتیجه‌گیری :**
  - **واریانس معیار خوبی برای روش‌هایی مثل PCA است که هدفشان حفظ اطلاعات خطی و واریانس کل است. اما برای روش‌هایی مثل t-SNE که ساختار محلی را اولویت می‌دهند، معیارهای دیگری (مثل حفظ فاصله‌های محلی) مهم‌ترند.**
  - **در این تمرین، نمودار واریانس برای PCA و Kernel PCA مفید بود، اما برایt-SNE ، مصورسازی مستقیم (مثل جداسازی خوشه‌ها) معیار بهتری بود.**

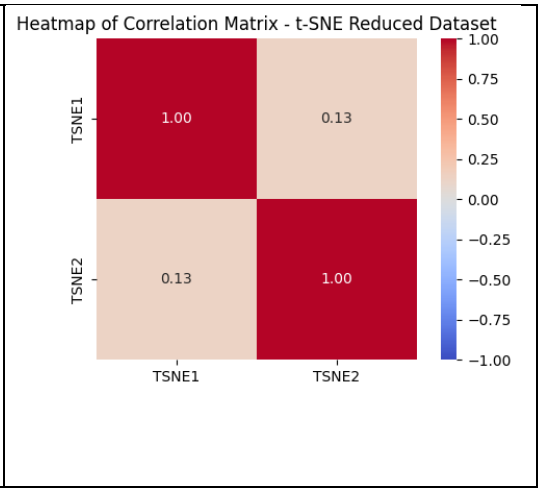
تحلیل مقایسه Heatmap اولیه و Heatmap پس از کاهش ابعاد

مقایسه Heatmap اولیه (Raw Dataset) با Heatmap پس از کاهش ابعادPCA) ، Kernel PCA، t-SNE)









تحلیل و مقایسه Heatmap ماتریس‌های همبستگی

۱ .مقایسه Heatmap اولیه باHeatmap های پس از کاهش ابعاد

- **Heatmapاولیه (Clean Dataset):**
  - این ماتریس همبستگی داده‌های اولیه(fnlwgt ، education-num، capital-gain، capital-loss، hours-per-week)را نشان می‌دهد.
  - مقادیر همبستگی بین ویژگی‌ها پایین است) بیشترین مقدار ۰.۱۴ بین education-num و-hours per-week).ویژگی fnlwgt با سایرین همبستگی بسیار ضعیف (نزدیک به صفر) دارد.
  - برخی همبستگی‌های ضعیف وجود دارد، مثل ۰.۱۲۵ بین education-num و capital-gain یا -۰.۰۳۱ بین capital-gain و.capital-loss
- **Heatmapپس از کاهش ابعاد :**

۱. **PCA Reduced Dataset:**

- ماتریس همبستگی بین مؤلفه‌های PC1 و PC2 نشان می‌دهد که همبستگی بین این دو مؤلفه صفر است (۰.۰۰). این نتیجه مورد انتظار است، زیرا PCA مؤلفه‌ها را به گونه‌ای تولید می‌کند که مستقل (uncorrelated)باشند.

۲. **Kernel PCA Reduced Dataset:**

- مشابهPCA ، همبستگی بین KPC1 و KPC2 نیز صفر است) ۰.۰۰ Kernel PCA .(نیز مؤلفه‌ها را مستقل تولید می‌کند، هرچند این استقلال در فضای غیرخطی تعریف می‌شود.

۳. **t-SNE Reduced Dataset:**

- برخلاف PCA وKernel PCA ، همبستگی بین TSNE1 و TSNE2 صفر نیست (۰.۱۳). این نشان می‌دهد که t-SNE روی استقلال مؤلفه‌ها تمرکز ندارد، بلکه هدفش حفظ ساختار محلی داده‌هاست.

۲ .تحلیل تغییرات ساختار ماتریس، استقلال ویژگی‌ها، و حذف هم‌پایگی‌ها

- **آیا ساختار ماتریس تغییر کرده؟**

- **PCA و Kernel PCA:** ساختار ماتریس به طور کامل تغییر کرده است. در داده‌های اولیه، ویژگی‌ها همبستگی‌های ضعیف (مثل ۰.۱۴) داشتند، اما پس از کاهش ابعاد، مؤلفه‌ها کاملاً مستقل شده‌اند (همبستگی صفر). این به دلیل ماهیت PCA است که مؤلفه‌ها را متعامد (orthogonal) می‌سازد.
- **t-SNE:** ساختار ماتریس تغییر کرده، اما مؤلفه‌ها همچنان همبستگی دارند) ۰.۱۳ t-SNE .(روی حفظ روابط محلی تمرکز دارد، نه استقلال مؤلفه‌ها، بنابراین ساختار همبستگی متفاوتی ایجاد می‌کند.

- **آیا ویژگی‌ها مستقل‌تر شده‌اند؟**
  - **PCA و Kernel PCA:** بله، کاملاً مستقل شده‌اند (همبستگی صفر). این نشان می‌دهد که این روش‌ها توانسته‌اند هرگونه همبستگی خطی یا غیرخطی) در مورد (Kernel PCA را حذف کنند.
  - **t-SNE:** خیر، مؤلفه‌ها همچنان همبستگی دارند) ۰.۱۳ t-SNE .(روی استقلال ویژگی‌ها تمرکز ندارد، بلکه ساختار محلی را حفظ می‌کند.
- **کاهش بعد چقدر توانسته هم‌پایگی‌های غیرضروری را حذف کند؟**
  - **PCA و Kernel PCA:** این روش‌ها تمام همبستگی‌ها را حذف کرده‌اند (همبستگی صفر بین مؤلفه‌ها). در داده‌های اولیه، همبستگی‌های ضعیف) مثل ۰.۱۴ بین education-num و (hours-per-week وجود داشت که ممکن است غیرضروری باشند. این روش‌ها با تولید مؤلفه‌های مستقل، این هم‌پایگی‌ها را کاملاً حذف کرده‌اند.
  - **t-SNE:** هم‌پایگی‌ها حذف نشده‌اند (همبستگی ۰.۱۳ باقی مانده است) t-SNE .برای حذف همبستگی طراحی نشده، بلکه برای نمایش ساختار داده‌ها به کار می‌رود.

۳ .نتیجه‌گیری در یک پاراگراف

کاهش ابعاد تأثیرات مثبتی فراتر از کاهش تعداد ویژگی‌ها دارد؛ این فرایند می‌تواند همبستگی‌های غیرضروری را حذف کند)مانند PCA و Kernel PCA که مؤلفه‌ها را کاملاً مستقل کردند(، داده‌ها را برای مصورسازی ساده‌تر کند) مانند t-SNE که خوشه‌ها را بهتر نشان داد(، و کارایی مدل‌های یادگیری ماشین را با کاهش پیچیدگی افزایش دهد. برای این مجموعه‌داده، PCA مناسب‌تر به نظر می‌رسد، زیرا همبستگی‌های اولیه پایین بودند و PCA توانست با حفظ واریانس (حدود ۶۰-۷۰%) و تولید مؤلفه‌های مستقل، ساختار داده‌ها را ساده‌تر کند، در حالی که t-SNE همبستگی‌ها را حفظ کرد و Kernel PCA پیچیدگی بیشتری اضافه کرد. ویژگی‌های education-num و hours-per-week در داده اولیه اهمیت بیشتری داشتند، زیرا بالاترین همبستگی (۰.۱۴) را نشان دادند و در مصورسازی‌ها) مثل ۲ (D PCA به تفکیک کلاس‌ها کمک کردند. کاهش ابعاد همچنین می‌تواند به شناسایی ویژگی‌های کم‌اهمیت) مثل fnlwgt با همبستگی نزدیک به صفر (کمک کند و از بیش‌برازش (overfitting)در مدل‌ها جلوگیری کند.