دانشگاه صنعتي اميركبير
(پلی‌تکنیک تهران)
دانشکده ریاضی و علوم کامپیوتر

گزارش پروژه درس داده کاوی محاسباتی پروژه 2

استخراج ویژگی مبتی بر استقلال خطی و تحلیل تاثیر آن بر فرایند بهینه سازی توابع زیان

نگارش
هومن ذوالفقاری

استاد راهنما
دکتر مهدی قطعی

تدریس یار
مهندس بهنام یوسفی مهر

آبان 1404

صفحه فرم ارزیابی و تصویب پایان‌نامه- فرم تأیید اعضاي کمیته دفاع

در این صفحه (هر سه مقطع تحصیلي) باید تصویر فرم ارزیابي یا تأیید و تصویب پایان-نامه/رساله موسوم به فرم کمیته دفاع براي مقاطع کارشناسی‌ارشد و دکتری و تصویر فرم تصویب براي مقطع کارشناسي، موجود در پرونده آموزشی را قرار دهند.

نکات مهم:

- ✓ نگارش پایان‌نامه/رساله باید به زبان فارسی و بر اساس آخرین نسخه دستورالعمل و راهنمای تدوین پایان‌نامه‌های دانشگاه صنعتی امیرکبیر باشد. (دستورالعمل و راهنمای حاضر)؛
- ✓ تحویل پایان‌نامه به زبان انگلیسي، براي دانشجویان بین‌الملل با شرایط دستورالعمل حاضر بلامانع است و داشتن صفحه عنوان فارسي به همراه چکیده مبسوط فارسي، 30 صفحه براي پایان‌نامه کارشناسی‌ارشد و50 صفحه براي رساله دکتری در ابتداي آن الزامي است؛
- ✓ دریافت پایان‌نامه کارشناسي،کارشناسي‌ارشد و رساله دکتری، بصورت نسخه دیجیتال مطابق راهنمای وبسایت و دستورالعمل حاضر مي‌باشد؛
- ✓ درصورتي که یک عنوان پایان‌نامه داراي دو نویسنده است، فقط یکبار فایل و فرم اطلاعات آن با ذکر هر دو نویسنده بارگذاري و تکمیل گردد؛
- ✓ با توجه به اینکه در ورود2016یا بالاتر، احتمال تغییر ترتیب ذکر زیر فصل‌ها وجود دارد لطفا در انتها به شماره‌دهي زیر فصل‌ها توجه نمایید که بصورت صحیح باشد. از راست به چپ: شماره فصل-زیرفصل1-زیرفصل2-زیرفصل3 و ... .

به نام خدا

تاریخ:

تعهدنامه اصالت اثر

اینجانب هومن ذوالفقاری متعهد می‌شوم که مطالب مندرج در این پایان نامه حاصل کار پژوهشی اینجانب تحت نظارت و راهنمایی اساتید دانشگاه صنعتي اميركبير بوده و به دستاوردهاي ديگران كه در اين پژوهش از آنها استفاده شده است مطابق مقررات و روال متعارف ارجاع و در فهرست منابع و مآخذ ذكر گرديده است. اين پايان نامه قبلاً براي احراز هيچ مدرك هم‌سطح يا بالاتر ارائه نگرديده است.

در صورت اثبات تخلف در هر زمان، مدرك تحصيلي صادر شده توسط دانشگاه از درجه اعتبار ساقط بوده و دانشگاه حق پيگيري قانوني خواهد داشت.

كليه نتايج و حقوق حاصل از اين پايان نامه متعلق به دانشگاه صنعتي اميركبير مي‌باشد. هرگونه استفاده از نتايج علمي و عملي، واگذاري اطلاعات به ديگران يا چاپ و تكثير، نسخه‌برداري، ترجمه و اقتباس از اين پايان نامه بدون موافقت كتبي دانشگاه صنعتي اميركبير ممنوع است.
نقل مطالب با ذكر مآخذ بلامانع است.

در صفحه تعهدنامه اصالت اثر، در قسمت بالا سمت چپ، تاریخ دفاع خود را جایگزین تاریخ نوشته شده کنید.

همچنین در صفحه تعهدنامه اصالت اثر، در خط اول، نام و نام خانوادگی خود را به صورت کامل با نام و نام خانوادگی نمونه، جایگزین کنید. در انتهای متن تعهد، در قسمت امضا نیز باید نام و نام خانوادگی کامل خود را وارد نماید.

هومن ذوالفقاری

امضا

چکیده

چکیده باید جامع و بیان‌کننده خلاصه‌اي از اقدامات انجام‌شده باشد. در قسمت چکیده، چکیده پایان‌نامه خود را که حداکثر می‌تواند شامل 250 کلمه باشد، بنویسید. در آخر چکیده و در قسمت واژگان کلیدی، کلمات کلیدی خود را وارد کنید. کلمات کلیدی بین 3 تا 5 کلمه می‌توانند باشد که طبق فرمت باید با ویرگول از هم جدا شوند.

واژه‌های کلیدی:

داده کاوی محاسباتی، کاهش ابعاد، انتخاب فیچر

فهرست مطالب

صفحه

صفحه

فهرست اشکال

فهرست جداول

صفحه

فهرست علائم

علائم لاتين

h	ارتفاع
L	طول موج توربولانس
T	پريود توربولانس
U_0	سرعت تعادل وسيله پرنده
$u_{\mathrm{g}}^{\mathrm{B}}$	مولفه سرعت تندباد در راستاي محور طولي دستگاه مختصات بدني نسبت به اينرسي

علائم يوناني

$\Phi(\omega)$	چگالي طيفي قدرت توربولانس
σ	شدت توربولانس
ω	بسامد توربولانس
Ω	بسامد فاصله‌اي

بالانويس‌ها

B	دستگاه مختصات بدني

زيرنويس‌ها

g	تندباد (گاست)

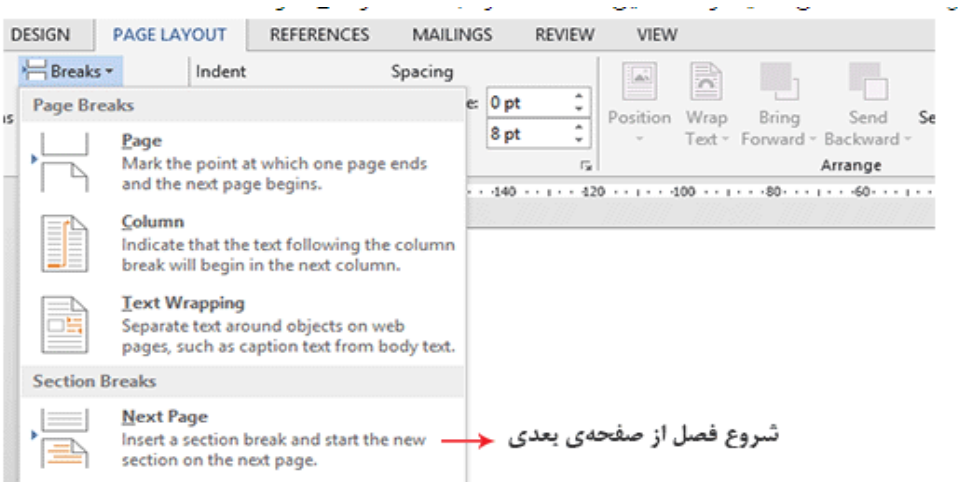
فصل اول
مقدمه (دستور العمل)

مقدمه

1- پایان‌نامه نمونه برای 5فصل طراحی شده است، چنان‌چه تعداد فصل‌های پایان‌نامه شما، کمتراز پنج فصل است، فصول اضافه را پاک کنید.

2- اگر تعداد فصل‌ها، بیشتر از پنج فصل باشد، برای اضافه کردن یک فصل جدید، باید قسمت(Section) جدیدی ایجاد کنید. برای ایجاد یک قسمت جدید با تنظیمات متفاوت نسبت به قسمت قبل در یک فایل ورد، کافیست مکان نما را در جایی که باید قسمت جدید آغاز شود (مثلا پس از اتمام فصل5) قرار دهید. سپس از تب Page Layout و از گروه Page Setup، روی دکمه‌ی بازشونده‌ی Breaks کلیک کنید. همان طور که می‌بینید، چهار نوع Section Break وجود دارد:

- Next Page : فصل جدید از صفحه‌ی بعد شروع می‌شود.
- Continuous : بدون شکست صفحه، فصل جدید در ادامه‌ی فصل قبلی شروع می‌شود.
- Even Page : فصل جدید از نخستین صفحه‌ی زوج بعدی آغاز می‌شود.
- Odd Page : فصل جدید از نخستین صفحه‌ی فرد بعدی آغاز می‌شود.



شکل1: نحوه ساخت قسمت جدید

مطابق شکل (1) گزینه Next Page را انتخاب کنید. در ادامه برای تنظیم سربرگ فصل جدید، در قسمت سربرگ صفحه دوبار کلیک کنید تا قابلیت ویرایش آن فعال شود و بخش Header and Footer Tools به تب‌ها اضافه شود. در تب Design و در گروه Navigation، دکمه‌ی Link to

previous را پیدا کنید. همان طور که می‌بینید، این دکمه به طور پیش‌فرض روشن است. روی آن کلیک کنید تا دکمه خاموش شود و ارتباط این فصل با فصل پیشین قطع شود. اکنون اگر سربرگ قسمت فعلی را ویرایش کنید، سربرگ قسمت قبلی تغییر نخواهد کرد. برای اینکار باید گزینه Link to Previous را غیرفعال کرد.

برای سرفصل ها مطابق جدول « جدول 4-1 قلم‌هاي فارسي» اقدام کنید و یکی از سرفصل مورد نظر را در این قالب کپی کنید و بعد جایگزین (paste) کنید، چنانچه بعد از جایگزین کردن شماره سرفصل به هر دلیلی دچار بهم ریختگی شد از راست به چپ، عدد چهارم را تغییر داده و سپس متن آن را ویرایش کنید.

- در هرقسمتی از پایان‌نامه که نیاز به جدول دارید، مراحل زیر را انجام دهید:

1- یکی از جدول‌های موجود در متن نمونه همراه با عنوان آن به طور کامل انتخاب کنید.

2- با کلید ترکیبی ctrl+c آن جدول را کپی کنید.

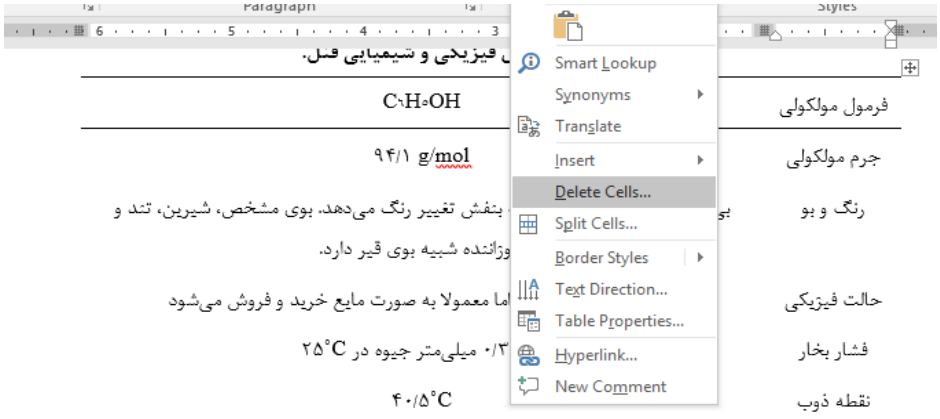
3- سپس به محل مورد نظر رفته و درآنجا با کلید ترکیبیctrl+v جدول را جایگزین(paste) کنید.

4- داده‌های مورد نظر خود را در جدول وارد کنید.

5- اگر در جدول انتخابی نیاز به اضافه کردن سطر یا ستون است، بر روی جدول کلیک راست کرده و از گزینه Insert، عملیات مورد نظر را انتخاب کنید(شکل 2).

شکل 2: اضافه کردن سطر یا ستون در جدول

6- اگر تعداد سطر و ستون جدول شما، کمتر از جدول انتخابی است، در این حالت نیاز به حذف تعدادی سطر یا ستون دارید. برای اینکار بر روی جدول کلیک راست کرده و با انتخاب گزینهDeletecells...، پنجره زیر باز می‌شود (شکل3)، برای حذف سطر اضافی گزینه سوم و برای حذف ستون اضافی گزینه چهارم را انتخاب کنید.



شکل 3: حذف سطر یا ستون در جدول

اگر به هر دلیلی دچار بهم ریختگی شدید مطابق « جدول 4-2 قلم‌هاي فارسي» از سبك(Style) مورد اشاره استفاده و **در آخر هم برای تهیه فهرست جداول ، از استایل TABLE TITLE فهرست گیری کنید.**

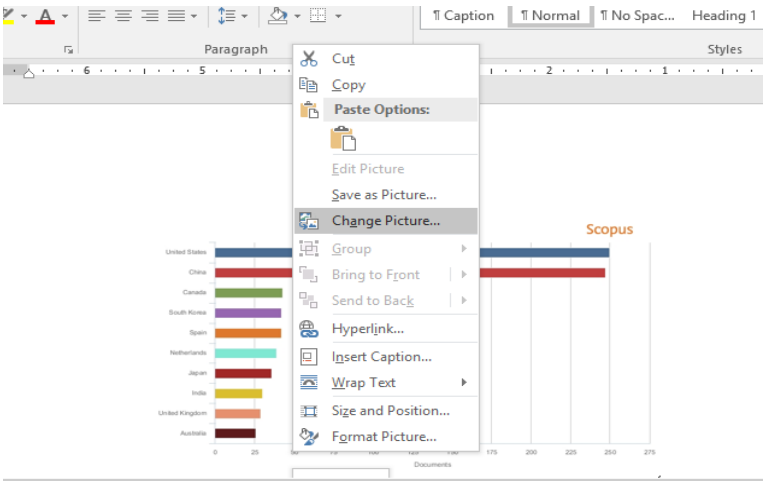
- برای قرار دادن عکس در پایان نامه، باید مراحل زیر را به ترتیب انجام دهید:

1- یکی از عکس‌های موجود در متن پایان‌نامه نمونه همراه با عنوان آن به طور کامل انتخاب کنید.

2- با کلید ترکیبی ctrl+c آن عکس را کپی کنید.

3- سپس به محل مورد نظر رفته و در آنجا با کلید ترکیبی ctrl+v عکس را جایگزین کنید.

4- روی عکس کلیک راست کرده مطابق شکل 4، گزینهchange picture... انتخاب کنید.



شکل4: تغییر عکس در پایان‌نامه

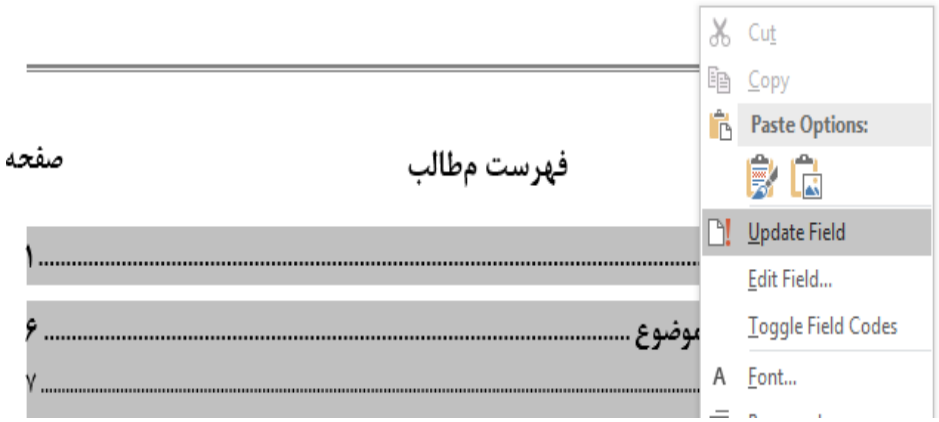
5- با انتخاب گزینه بالا، صفحه زیر باز می‌شود، از قسمت جستجو (Browse)، عکس مورد نظر را انتخاب کرده و جایگزین عکس فعلی می‌کنید.

6- متن موردنظر خود را جایگزین متن عکس انتخابی کنید. البته نباید در سایز و فونت آن تغییری دهید.

اگر به هر دلیلی دچار بهم ریختگی شدید مطابق « جدول 4-3 قلم‌هاي فارسي» از سبك مورد اشاره استفاده و **در آخر هم برای تهیه فهرست جداول ، از سبك PIC TITLE فهرست‌گیری کنید.**

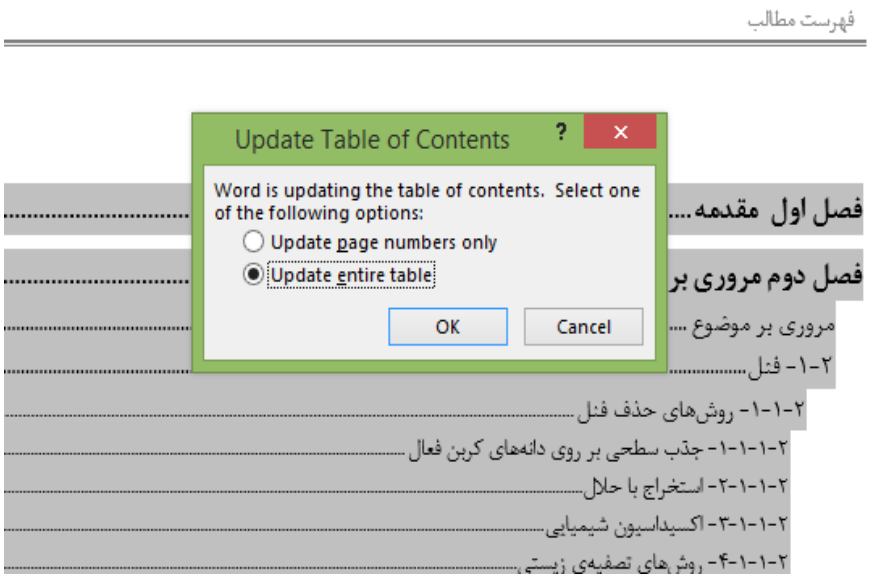
بعداز وارد و جایگزین کردن کامل مطالب خود در پایان‌نامه نمونه، به قسمت فهرست مطالب برگردید.

18- برای بروزرساني فهرست‌ها(مطالب، اشکال و جداول)، بر روی جدول فهرست مورد نظر کلیک راست کرده و گزینه Update Field را انتخاب کنید.(شکل 5).



شک5: انتخاب گزینه بروزرسانی فهرست

19- بعد از انتخاب این گزینه، پنجره زیر باز می‌شود(شکل6)، برای بروزرسانی کامل فهرست گزینه دوم را انتخاب کنید. فهرست جدید برای پایان‌نامه شما ساخته می‌شود.

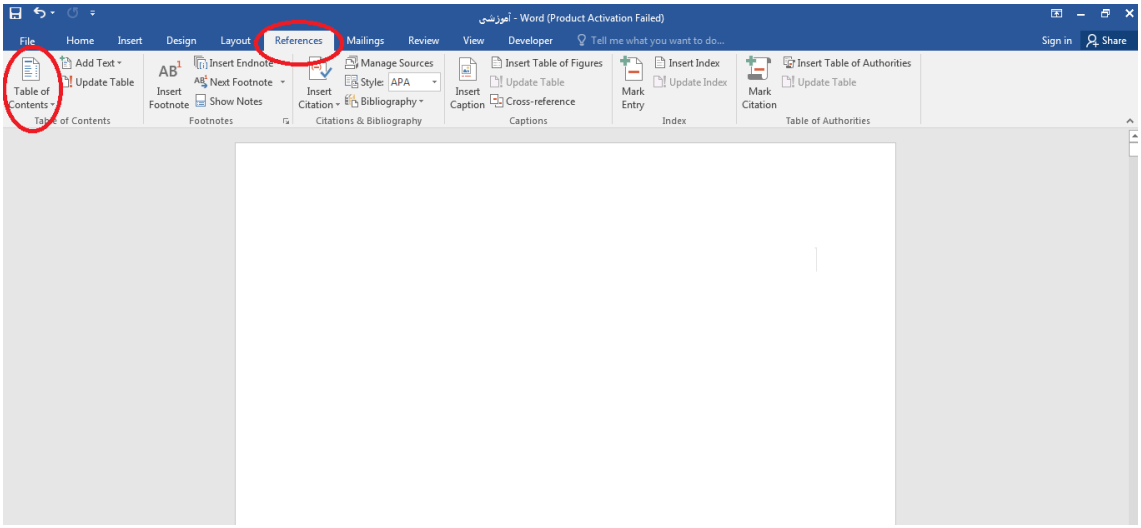


شکل6: بروز رسانی کامل فهرست

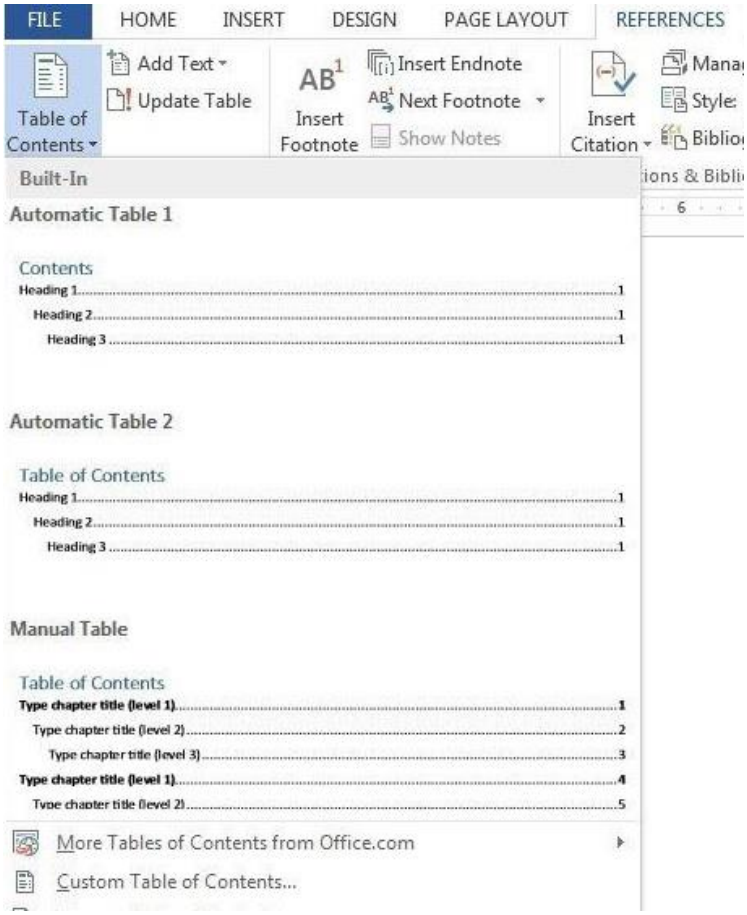
در آخر هم در خصوص فهرست گیری در صورتی که با جایگزین کردن فهرست دچار مشکل شدید از HEADING1 و HEADING2 و HEADING3 و HEDING4 مطابق دستورالعمل زیر فهرستگیری نمایید.

نحوه‌ی نوشتن فهرست مطالب

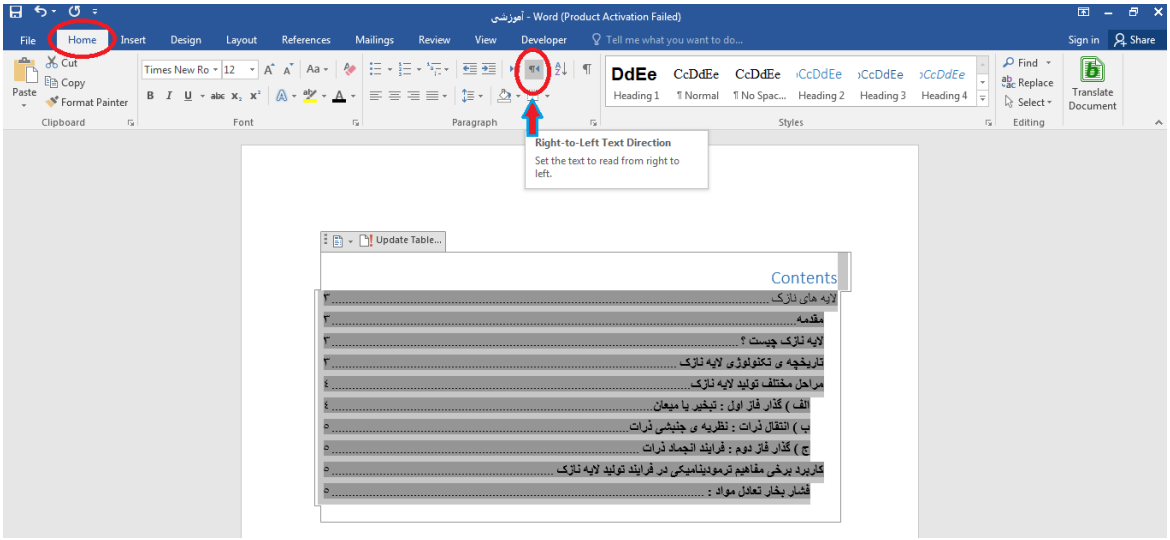
از سربرگ‌های موجود در صفحه‌ی ورد خود وارد سربرگ References شوید و بر روی گزینه ی Table of Contents کلیک کنید.



با کلیک بر روی Table of Contents نمونه حالت‌های پیش فرض فهرست‌بندی در ورد به صورت الگو برایتان به نمایش در می‌آید. که با انتخاب هر کدام از آنها فهرست شما به همان شکل به فایل وردتان اضافه می‌شود.



باید توجه داشت که فهرست ایجاد شده برای زبان فارسی مناسب نیست، کافی است تا کل فهرست را انتخاب کرده و سپس در تب Home بر روی گزینه‌ی Right-to-Left کلیک کنید تا فهرست از راست به چپ قرار گیرد.



حالا به ویرایش فهرست می‌رسیم، می‌توانیم عبارت Contents را پاک کرده و به جای آن عبارت "فهرست" را قرار دهیم و با استفاده از قسمت Font نوع فونت و اندازه‌ی فونت متن فهرست و عناوین را به دلخواه تنظیم کنیم.

فصل دوم
داده ها

داده ها

2-1- داده های کلاس بندی

مجموعه‌داده سرطان پستان ویسکانسین (تشخیصی):

ویژگی‌های مجموعه‌داده:

- **تعداد نمونه‌ها :** ۵۶۹
- **تعداد ویژگی‌ها :** ۳۰ ویژگی عددی پیش‌بین و یک ویژگی کلاسی

توضیحات ویژگی‌ها:

- شعاع (میانگین فاصله از مرکز تا نقاط روی محیط)
- بافت (انحراف معیار مقادیر خاکستری)
- محیط
- مساحت
- یکنواختی (تغییرات محلی در طول شعاع)
- فشردگی (محیط^2 / مساحت - ۱/۰)
- فرورفتگی (شدت بخش‌های فرورفته در مرز)
- نقاط فرورفته (تعداد بخش‌های فرورفته در مرز)
- تقارن
- بعد فراکتال ("تقریب خط ساحلی" - ۱)

کلاس‌ها:

- WDBC-Malignant (بدخیم)
- WDBC-Benign (خوش‌خیم)

توزیع کلاس‌ها:

- ۲۱۲ مورد بدخیم
- ۳۵۷ مورد خوش‌خیم

2-2-داده های رگرسیون

```
(20640, 8) (20640,)
   MedInc  HouseAge  AveRooms  AveBedrms  Population  AveOccup  Latitude  \
0  8.3252      41.0  6.984127   1.023810       322.0  2.555556     37.88
1  8.3014      21.0  6.238137   0.971880      2401.0  2.109842     37.86
2  7.2574      52.0  8.288136   1.073446       496.0  2.802260     37.85
3  5.6431      52.0  5.817352   1.073059       558.0  2.547945     37.85
4  3.8462      52.0  6.281853   1.081081       565.0  2.181467     37.85

   Longitude
0    -122.23
1    -122.22
2    -122.24
3    -122.25
4    -122.25
```

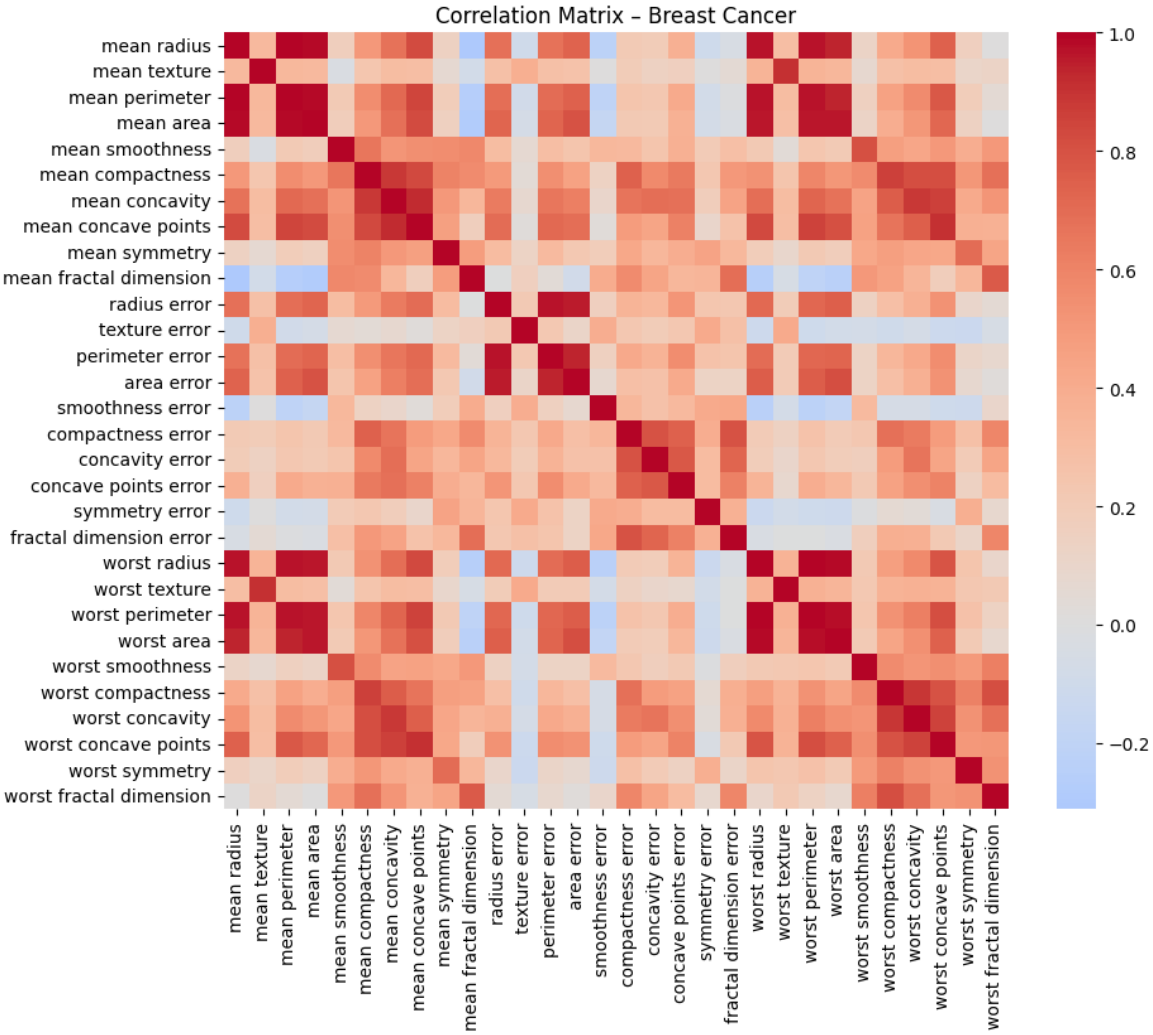
2-3-داده های خوشه بندی

```
(150, 4) (150,)
['setosa' 'versicolor' 'virginica']
```

فصل سوم
بررسی کورلیشن

بررسی کورلیشن

3-1- داده های طبقه بندی



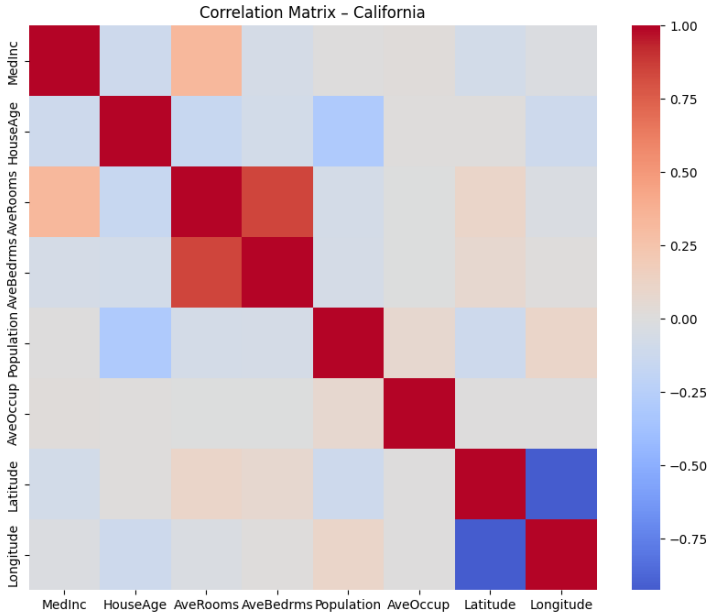
خلاصهٔ ویژگی‌های با همبستگی بالا در مجموعه‌داده سرطان پستان (ضریب همبستگی > 0.75)

- ویژگی‌های **شعاع، محیط و مساحت** (در حالت میانگین و بدترین مقدار) تقریباً هم‌پوشانی کامل دارند؛ ضریب همبستگی بین آن‌ها تا **۱.۰۰** می‌رسد.
- **نقاط فرورفته(concave points)** ، **فرورفتگی (concavity)** و **فشردگی (compactness)**نیز به‌طور قوی با یکدیگر مرتبطاند (تا حدود **۰.۹۲**).
- نسخه‌های **میانگین** و **بدترین** هر ویژگی (مثل بافت، یکنواختی، بعد فراکتال و...) معمولاً همبستگی بالایی دارند (بین **۰.۸ تا ۰.۹**).

- خطاهای (error) مربوط به **شعاع، محیط و مساحت** نیز بسیار همبسته‌اند (حدود **۰.۹۵**–**۰.۹۷**).
- در کل، مجموعه‌داده شامل چند **خوشهٔ بزرگ از ویژگی‌های همبسته** است که اطلاعات مشابهی ارائه می‌دهند:
  1. خوشهٔ اندازه(radius–perimeter–area)
  2. خوشهٔ شکل(compactness–concavity–concave points)
  3. خوشهٔ خطاها(radius error–perimeter error–area error)
  4. خوشهٔ نسخه‌های «میانگین» و «بدترین» هر ویژگی

نتیجه: بسیاری از ویژگی‌ها در این داده **همبستگی بسیار بالا** دارند، بنابراین در روش‌های کاهش بُعد) مثل PCA یا انتخاب ویژگی، می‌توان برخی از آن‌ها را حذف کرد بدون از دست دادن اطلاعات مهم.

2-3- داده های رگرسیون

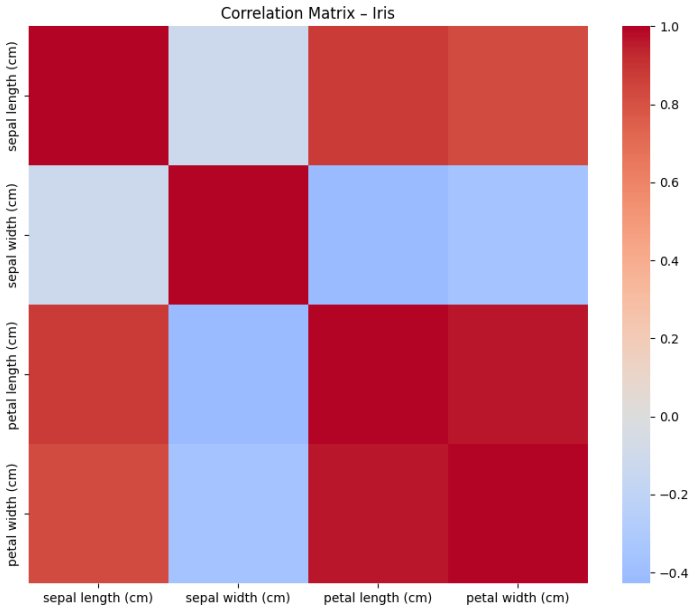


California – Highly Correlated Features (>0.75):

AveRooms ↔ AveBedrms: 0.85

Latitude ↔ Longitude: 0.92

3-3-1- داده های خوشه بندی



Iris – Highly Correlated Features (>0.75):

sepal length (cm) ↔ petal length (cm): 0.87

sepal length (cm) ↔ petal width (cm): 0.82

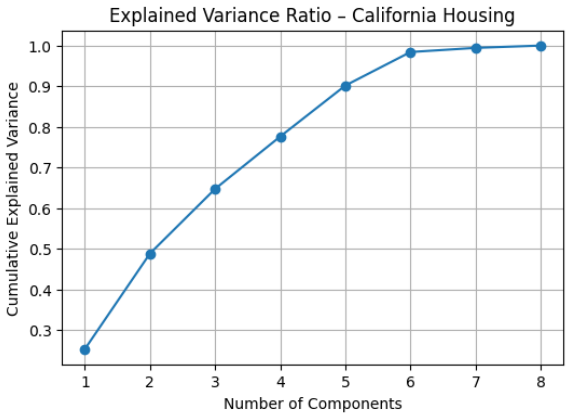
petal length (cm) ↔ petal width (cm): 0.96.

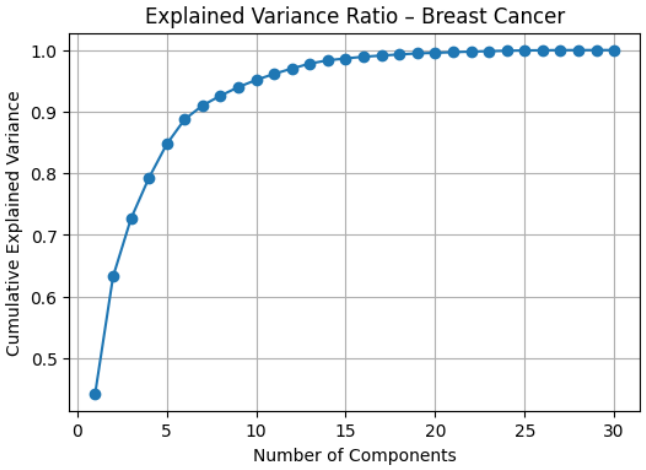
فصل چهارم
کاهش بعد و انتخاب فیچر

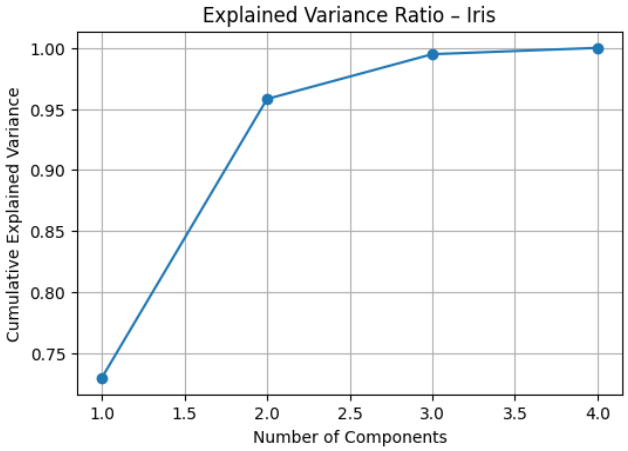
کاهش بعد و انتخاب فیچر

در تعريف سبك‌هاي مختلف اين دستورالعمل از قلم‌هاي، B Nazanin و Times New Roman استفاده شده‌است كه خصوصيات كامل آنها در بخش‌هاي بعدي تشريح مي‌گردد.

4-1-کاهش بعد







- **خلاصه نتایج تحلیل مؤلفه‌ها و خطاها در سه مجموعه‌داده:**

- **Breast Cancer**
- **تعداد مؤلفه‌ها برای ۹۵٪ واریانس: ۱۰**
- **میانگین خطای بازسازی (PCA MSE): ۰٫۰۴۸۴**

- **میـــــــــــانگین قـــــــــــدرمطلق همبســـــــــــتگی ویژگی‌هـــــــــــا:**
  PCA = ۰/۱۰۰ | ICA = ۰/۱۰۰ | SVD = ۰/۱۰۰
- *توضیح* :این مجموعه‌داده پیچیده‌تر است و برای پوشش ٪۹۵ واریـانس بـه ۱۰ مؤلفـه نیـاز دارد.
-
- **California Housing**
- **تعداد مؤلفه‌ها برای ٪۹۵ واریانس :**۶
- **PCA MSE:** ۰/۰۱۵۹

- **Mean|Corr|:**
  PCA = ۰/۱۶۷ | ICA = ۰/۱۶۷ | SVD = ۰/۱۶۷
- *توضیح* :ویژگی‌ها همبستگی نسبتاً بالاتری دارند اما با مؤلفه‌های کمتر می‌توان واریـانس را توضیح داد.

- **Iris**
- **تعداد مؤلفه‌ها برای ٪۹۵ واریانس :**۲
- **PCA MSE:** ۰/۰۴۱۹

- **Mean|Corr|:**
  PCA = ۰/۵۰۰ | ICA = ۰/۵۰۰ | SVD = ۰/۵۰۰
- *توضیح* :دادهٔ آیریس کم‌بعد است و تنها با دو مؤلفه می‌توان تقریبـاً تمـام واریـانس را پوشـش داد. همبستگی ویژگی‌ها در این مجموعه بالاتر است.

- **جمع‌بندی کلی:**
- مجموعهٔ **Breast Cancer** بیشترین تعداد مؤلفه‌ها را برای حفظ واریانس نیاز دارد.
- در **California Housing** همبستگی ویژگی‌ها متوسط است.
- مجموعهٔ **Iris** با ابعاد کمتر و همبستگی بالا ساده‌تر فشرده می‌شود.

2-4- انتخاب فیچیر

سرطان پستان (دسته‌بندی)

SelectKBest: میـانگین شـعاع، محـیط، مسـاحت، فرورفتگـی، نقـاط فرورفتـه، و همچنـین بـدترین شـــــــعاع، محـــــــیط، مســـــــاحت، فرورفتگـــــــی و نقـــــــاط فرورفتـــــــه.

RFE: نقاط فرورفتهٔ میانگین، خطای شعاع، خطای مساحت، خطای فشردگی، و همچنین بدترین شعاع، بافت، محیط، مساحت، فرورفتگی و نقاط فرورفته.
توضیح: هر دو روش بیشتر بر ویژگی‌های مربوط به شکل و اندازه توده‌ها تمرکز دارند.

مسکن کالیفرنیا (رگرسیون)

SelectKBest: درآمد متوسط، سن ساختمان، میانگین تعداد اتاق‌ها، میانگین تعداد اتاق‌خواب‌ها، عرض جغرافیایی.
RFE: درآمد متوسط، میانگین تعداد اتاق‌ها، میانگین تعداد اتاق‌خواب‌ها، عرض و طول جغرافیایی.
توضیح: هر دو روش نشان می‌دهند که درآمد و ویژگی‌های مکانی مهم‌ترین عوامل در تعیین قیمت مسکن هستند.

گل آیریس (دسته‌بندی نظارت‌شده)

SelectKBest: طول کاسبرگ، طول گلبرگ، عرض گلبرگ.
RFE: عرض کاسبرگ، طول گلبرگ، عرض گلبرگ.
توضیح: هر دو روش نشان می‌دهند که ویژگی‌های مربوط به گلبرگ‌ها بیشترین نقش را در تفکیک گونه‌های آیریس دارند.

گل آیریس (خوشه‌بندی بدون نظارت)

VarianceThreshold: تمام ویژگی‌ها انتخاب شده‌اند؛ طول و عرض کاسبرگ، طول و عرض گلبرگ.
توضیح: در روش بدون نظارت، همهٔ ویژگی‌ها تنوع کافی دارند و حذف هیچ‌کدام ضروری نیست.

فصل پنجم
تحلیل بهینه سازی

تحلیل بهینه سازی

5-1- رگرسیون تحلیلی

خلاصه نتایج مدل‌ها برای مجموعه‌داده مسکن کالیفرنیا (20640 نمونه، 8 ویژگی):

مدل ۱: داده‌های اصلی

- مهم‌ترین ضرایب:

Latitude: -0.8999
Longitude: -0.8705
MedInc: 0.8296
AveBedrms: 0.3057
AveRooms: -0.2655
HouseAge: 0.1188
AveOccup: -0.0393
Population: -0.0045

مدل ۲: داده‌های PCA

- مهم‌ترین مؤلفه‌ها:

PC4: 0.7414
PC6: 0.3202
PC2: 0.1363
PC3: 0.0398
PC1: 0.0282
PC5: 0.0039

مدل ۳a: داده‌های SelectKBest

- ویژگی‌های انتخاب‌شده: MedInc، HouseAge، AveRooms، AveBedrms، Latitude
- ضرایب: [0.9998, 0.2012, -0.4888, 0.4428, -0.0677]

مدل ۳b: داده‌های RFE

- ویژگی‌های انتخاب‌شده: MedInc، AveRooms، AveBedrms، Latitude، Longitude
- ضرایب: [0.8095, -0.2735, 0.3096, -0.9842, -0.9625]

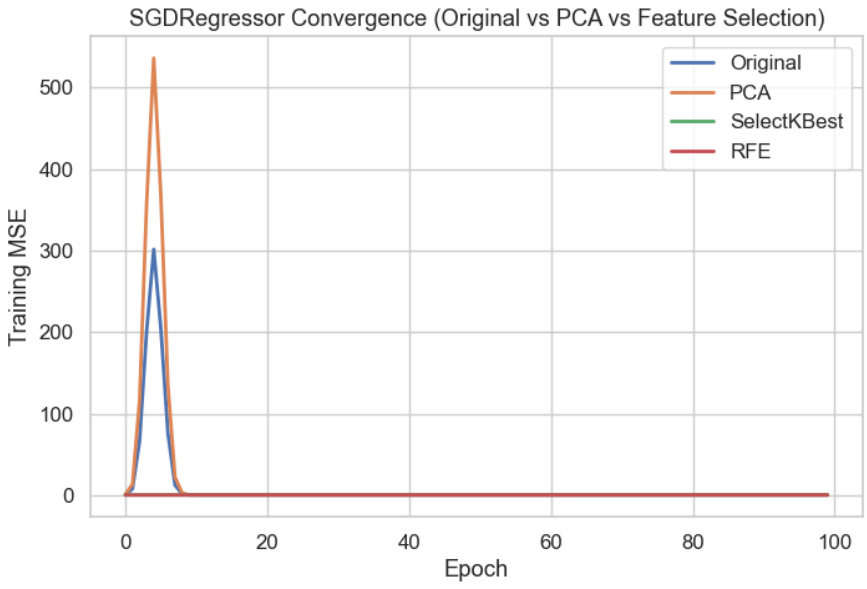
خلاصه ثبات ضرایب:

Dataset	Mean(|coef|)	Std(coef)
Original	0.4167	0.5420
PCA	0.2116	0.2596
SelectKBest	0.4400	0.4986
RFE	0.6679	0.7041

نتیجه:

- مدل RFE بیشترین میانگین و پراکندگی ضرایب را دارد، یعنی تأثیر ویژگی‌ها قوی‌تر و متغیرتر است.
- مدل PCA کمترین مقدار میانگین و ثبات ضرایب را دارد، که نشان‌دهنده کاهش قدرت تفکیک ویژگی‌ها پس از کاهش بُعد است.
- مدل‌های SelectKBest و داده‌های اصلی نسبتاً نزدیک به هم هستند اما SelectKBest کمی پایدارتر است.

5-2- رگرسیون با کاهش



خلاصه نتایج کاهش بعد و رگرسیون SGD برای مجموعه‌داده مسکن کالیفرنیا:

- **تعداد مؤلفه‌ها و ویژگی‌های انتخاب‌شده:**
  - مؤلفهPCA: 6

- SelectKBest: MedInc، HouseAge، AveRooms، AveBedrms، Latitude
- RFE: MedInc، Population، AveOccup، Latitude، Longitude

نتایج همگرایی رگرسیون:SGD

Dataset	Final MSE	Convergence Epoch
Original	0.5271	46
PCA	0.6607	28
SelectKBest	0.6169	3
RFE	0.5508	34

توضیح:

- داده‌های اصلی بهترین MSE نهایی را دارند اما به بیشترین تعداد اپک نیاز داشتند تا همگرا شوند.
- PCAبیشترین MSE و کمترین تعداد اپک همگرایی را نشان می‌دهد، که کاهش بعد باعث افت دقت شده است.
- SelectKBestبا تعداد کمی از اپک سریع همگرا شده اما MSE نسبت به داده‌های اصلی کمی بالاتر است.
- RFEتعادلی بین سرعت همگرایی و دقت نهایی ارائه می‌دهد.

5-3- خوشه بندی

خلاصه نتایج خوشه‌بندی KMeans روی مجموعه‌داده آیریس:

- **ویژگی‌های انتخاب‌شده با VarianceThreshold:**
  طول و عرض کاسبرگ، طول و عرض گلبرگ

نتایج خوشه‌بندی:KMeans

Dataset	Inertia	Silhouette	Iterations	Time (s)
Original	139.8205	0.4599	4	0.0144
PCA (2D)	115.0208	0.5092	4	0.0140

Dataset	Inertia	Silhouette	Iterations	Time (s)
SelectKBest	139.8205	0.4599	4	0.0180

تفسیر نتایج خوشه‌بندی:

- Inertiaکمتر :خوشه‌های فشرده‌تر :هرچه مقدار Inertia کمتر باشد، نمونه‌ها به مرکز خوشه‌های خود نزدیک‌تر هستند و خوشه‌ها متراکم‌ترند.
- Silhouetteبالاتر :جدایی بهتر خوشه‌ها :مقادیر بالاتر نشان می‌دهند که نمونه‌ها به خوشهٔ خود نزدیک‌تر و از خوشه‌های دیگر دورتر هستند، یعنی تفکیک خوشه‌ها بهتر است.
- تعداد اپک کمتر / زمان کمتر :همگرایی سریع‌تر :الگوریتم سریع‌تر به نقطهٔ بهینه می‌رسد و بهینه‌سازی مؤثرتر انجام شده است.

توضیح:

- کاهش بعد با PCA (دو مؤلفه) باعث کاهش Inertia و افزایش Silhouette شده است، یعنی خوشه‌ها تفکیک بهتری دارند.
- استفاده از ویژگی‌های SelectKBest عملکرد مشابه داده‌های اصلی را نشان می‌دهد، اما زمان اجرا کمی بیشتر است.

5-4- کلاس بندی نزدیک ترین همسایه

خلاصه نتایج KNN روی مجموعه‌داده سرطان پستان:

- **ویژگی‌های انتخاب‌شده با SelectKBest:**
میانگین محیط، میانگین نقاط فرورفته، بدترین شعاع، بدترین محیط، بدترین نقاط فرورفته

نتایج دسته‌بندی:KNN

Dataset	Accuracy	Prediction Time (s)	Train Time (s)
Original	0.95614	0.004865	0.000000
PCA	0.95614	0.002785	0.001009
SelectKBest	0.95614	0.000999	0.001000
RFE	0.95614	0.002000	0.000000

تفسیر:

- **دقت بالاتر → تعمیم بهتر مدل**
- **زمان پیش‌بینی کمتر → اجرای سریع‌تر مدل مبتنی بر نمونه**
- PCAبا کاهش بعد، معمولاً فاصله‌ها را بهتر نگه می‌دارد و سرعت KNN را افزایش می‌دهد.
- SelectKBestو RFE با حفظ ویژگی‌های تمایزی مهم، می‌توانند همان دقت یا حتی دقت بالاتر را با هزینه محاسباتی کمتر ارائه دهند.

نکته مهم: روش knn با توجه به وابتگی اش به فاصله ها در ابعاد بالا ضعیف عمل میکند که این مربوط به رفتار فاصله در هندسه ابعاد بالا است.

5-5- کلاس بندی جنگل تصادفی

خلاصه نتایج Random Forest روی مجموعه‌داده سرطان پستان:

- **ویژگی‌های انتخاب‌شده با SelectKBest:**
میانگین محیط، میانگین نقاط فرورفته، بدترین شعاع، بدترین محیط، بدترین نقاط فرورفته

نتایج دسته‌بندی:Random Forest

Dataset	Accuracy	Train Time (s)	Prediction Time (s)
Original	0.9561	0.1103	0.0145
PCA	0.9211	0.1215	0.0151
SelectKBest	0.9386	0.1469	0.0293
RFE	0.9474	0.2031	0.0260

تفسیر:

- Random Forestنسبت به مقیاس ویژگی‌ها مقاوم است و تحت تأثیر همخطی زیاد قرار نمی‌گیرد.
- استفاده از PCA گاهی باعث کاهش دقت می‌شود، زیرا ترکیب مؤلفه‌ها معنی اصلی ویژگی‌ها را مخلوط می‌کند.

- انتخاب ویژگی SelectKBest) و (RFE معمولاً دقت مشابه دادهٔ اصلی را ارائـه می‌دهـد و می‌تواند زمان آموزش را کاهش دهد یا تعادل بین زمان و دقت را تغییر دهد.
- مقایسه PCA با داده‌های اصلی نشان می‌دهد که کاهش بعد ممکـن اسـت زمـان آمـوزش را کمی افزایش دهد و دقت را کاهش دهد.

.

.

فصل هفتم
تحليل ها

تحلیل ها

5-2- گزارش نتایج:

۱ .رگرسیون(California Housing)

- شکل داده‌ها پس از پیش‌پردازش و کاهش بعد:
  - داده اصلی: 20640 نمونه × 8 ویژگی
  - PCA: 20640 × 6مؤلفه
  - SelectKBest: 20640 × 5ویژگی
  - RFE: 20640 × 5ویژگی
- نمودار همگرایی SGDRegressor نشان می‌دهد که کاهش بعد با PCA سرعت همگرایی را افزایش داده اما خطای نهایی MSE نسبت به داده اصلی کمی بیشتر است.
- SelectKBestو RFE عملکرد نزدیک به داده اصلی دارند و با تعداد کمی ویژگی مدل می‌توان دقت قابل قبول داشت.

۲ .خوشه‌بندی(Iris Dataset)

- KMeansروی سه فضای ویژگی انجام شد: داده اصلی، PCA(2 مؤلفه) و SelectKBest (2)ویژگی انتخاب‌شده.(
- نتایج:

Dataset	Inertia	Silhouette	Iterations	Time(s)
Original	139.8205	0.4599	4	0.0161
PCA	115.0208	0.5092	4	0.0120
SelectKBest	18.0270	0.6741	6	0.0133

- کاهش بعد با PCA باعث کاهش Inertia و افزایش Silhouette شده است، یعنی خوشه‌ها فشرده‌تر و جداسازی آنها بهتر است.

- SelectKBestبیشترین Silhouette و کمترین Inertia را نشـان داد، امـا تعـداد Iteration کمی بیشتر شد.

۳ .دسته‌بندی(Breast Cancer Dataset)

- **KNN:**

Dataset	Accuracy	PredictTime(s)	TrainTime(s)
Original	0.95614	0.00202	0.00000
PCA	0.95614	0.00100	0.00124
SelectKBest	0.95614	0.00122	0.00100
RFE	0.95614	0.00099	0.00103

- **Random Forest:**

Dataset	Accuracy	PredictTime(s)	TrainTime(s)
Original	0.95614	0.01572	0.12862
PCA	0.92105	0.01472	0.17456
SelectKBest	0.93860	0.02866	0.16110
RFE	0.94737	0.02882	0.12167

- KNNدقت یکسانی در تمام فضاهای ویژگـی نشـان داد، امـا زمـان پیش‌بینـی در PCA و RFEکمی کاهش یافت.
- Random Forestنسبت به PCA حسـاس‌تر اسـت و اسـتفاده از PCA باعـث کـاهش دقـت شد، در حالی که SelectKBest و RFE دقت بالا و زمان آموزش مناسب ارائه کردند.

6-2- تحلیل نتایج و نمودارها

- **۱ .تأثیر استقلال خطی ناشی از PCA بر پایداری ضرایبLinearRegression**
- استفاده از PCA باعث شد که ویژگی‌ها تقریباً مستقل (غیرهم‌خطی) شوند.
- جدول coef_summaryنشان می‌دهد که میانگین قدر مطلق ضرایب و انحراف معیار آنها در PCA کمتر از داده اصلی بود:
  - Original: Mean(|coef|) ≈ 0.417, Std(|coef|) ≈ 0.542

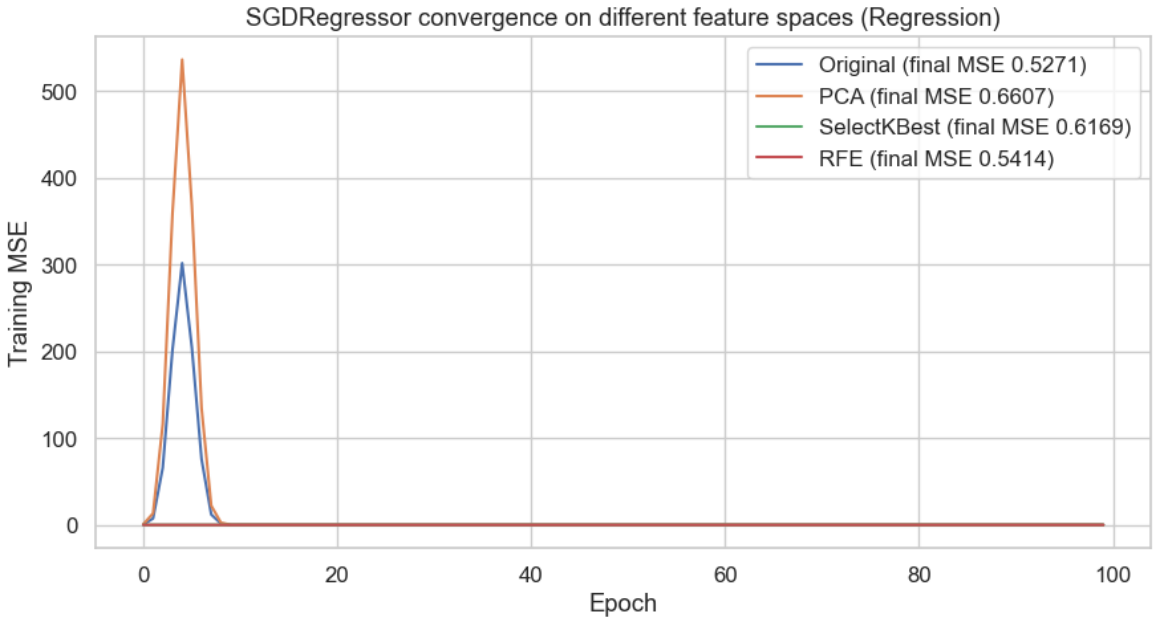
- PCA: Mean(|coef|) ≈ 0.212, Std(|coef|) ≈ 0.260
- این کاهش نشان می‌دهد که ضرایب LinearRegression در فضای PCA پایدارتر و کمتر تحت تأثیر نوسانات داده‌ها قرار گرفتند، زیرا هم‌خطی بین ویژگی‌ها کاهش یافته بود.
- **۲ .تأثیر استقلال خطی بر سرعت همگراییSGDRegressor**
- نمودار همگرایی SGDRegressor (sgd_convergence_regression.png) نشان می‌دهد که مدل روی داده PCA نسبت به داده اصلی سریع‌تر به خطای پایانی نزدیک شد:
  - Original: همگرایی حدود epoch46
  - PCA: همگرایی حدود epoch28
- دلیل: کاهش هم‌خطی باعث می‌شود که الگوریتم‌های گرادیان‌محور با گام‌های یکنواخت‌تر و بدون جهش‌های ناگهانی حرکت کنند، در نتیجه همگرایی سریع‌تر رخ می‌دهد.
- **۳ .تأثیر کاهش بعد با PCA بر عملکرد KMeans وKNN**
- **KMeans (Iris):**
  - PCA باعث کاهش Inertia و افزایش Silhouette شد(Inertia=115 vs 139 ، Silhouette=0.509 vs 0.460).
  - خوشه‌ها در فضای PCA فشرده‌تر و جداسازی آنها بهتر است.
  - دلیل: کاهش ابعاد حذف نویز و هم‌خطی بین ویژگی‌ها را انجام می‌دهد و فضای داده برای خوشه‌بندی مناسب‌تر می‌شود.
- **KNN (Breast Cancer):**
  - دقت ثابت ماند (Accuracy ≈ 0.956) اما زمان پیش‌بینی کمی کاهش یافت.
  - دلیل KNN :مبتنی بر فاصله است و کاهش ابعاد، فاصله‌ها را دقیق‌تر و محاسبات را سریع‌تر می‌کند.
- **۴ .عملکرد RandomForestClassifier روی داده‌های اصلی با هم‌خطی**
- دقت داده اصلی: 0.956
- دقت PCA کاهش یافت: 0.921
- SelectKBest/RFE: 0.939–0.947
- نتیجه RandomForest :نسبت به هم‌خطی مقاوم است و عملکرد خوبی روی داده اصلی دارد، زیرا درخت‌های تصمیم مستقل هستند و هر درخت فقط از یک زیرمجموعه از ویژگی‌ها استفاده می‌کند، بنابراین هم‌خطی بین ویژگی‌ها اثر منفی زیادی ندارد.

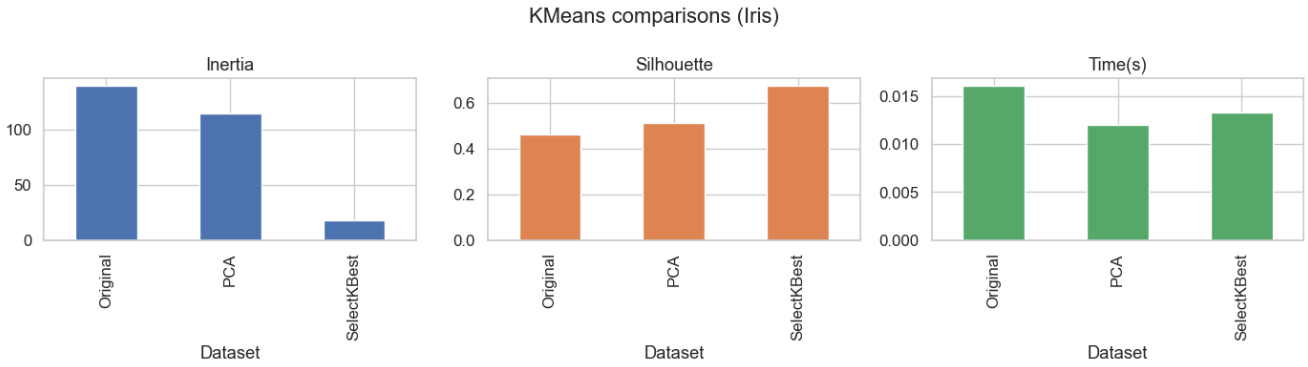
- **۵ .مقایسه استخراج ویژگی (PCA) با انتخاب ویژگی(SelectKBest/RFE)**
- PCA **:LinearRegression/SGDRegressor**باعث پایداری ضرایب و همگرایی سریع‌تر شد.
- PCA **:KMeans**باعث خوشه‌بندی بهتر شد، به ویژه کاهش نویز و همخطی.
- PCA **:KNN**کمی زمان پیش‌بینی را کاهش داد ولی دقت تغییر نکرد.
- SelectKBest/RFE **:RandomForest**عملکرد بهتری نسبت به PCA داشت (دقت بالاتر و زمان آموزش کمتر).
- نتیجه‌گیری کلی:
  - برای مدل‌های خطی و مبتنی بر فاصله، PCAمی‌تواند مفید باشد.
  - برای مدل‌های مبتنی بر درخت، انتخاب ویژگی (SelectKBest/RFE) بهتر است، زیرا حفظ ویژگی‌های اصلی برای تفکیک نمونه‌ها مهم است.

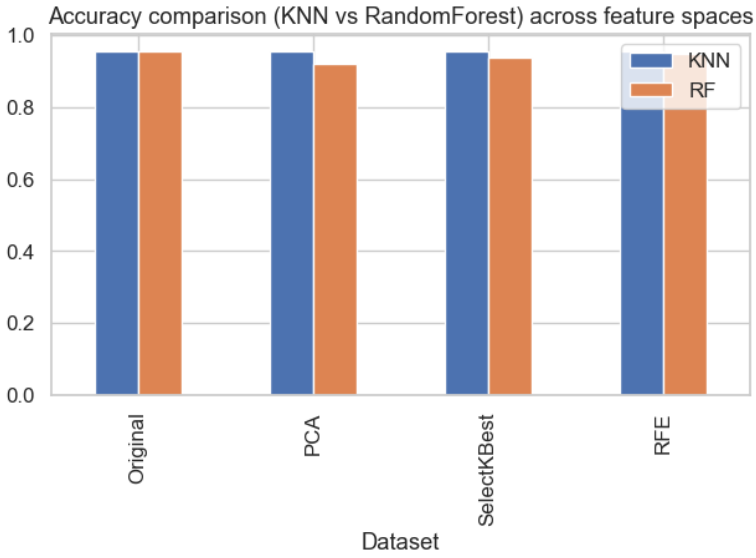
۴ .جمع‌بندی کلی

- کاهش بعد با PCA سرعت همگرایی و پردازش را افزایش می‌دهد، اما گاهی دقت مدل‌های پیچیده‌تر) مثل Random Forest) کاهش می‌یابد.
- انتخاب ویژگی با روش‌های SelectKBest و RFE تعادل خوبی بین دقت، زمان آموزش و تعداد ویژگی‌ها ایجاد می‌کند.
- نمودارها و جداول مربوطه در فایل‌های PNG و CSV ذخیره شده‌اند.

نمودار ها:


SGDRegressor convergence on different feature spaces (Regression)
Original (final MSE 0.5271)
PCA (final MSE 0.6607)
SelectKBest (final MSE 0.6169)
RFE (final MSE 0.5414)
Training MSE
Epoch
0
100
200
300
400
500
0
20
40
60
80
100


KMeans comparisons (Iris)
Inertia
Silhouette
Time(s)
Original
PCA
SelectKBest
Dataset


Accuracy comparison (KNN vs RandomForest) across feature spaces
KNN
RF
Original
PCA
SelectKBest
RFE
Dataset

فصل ششم
جمع‌بندي و نتيجه‌گيري

جمع‌بندي و نتيجه‌گيري

جمع‌بندی:

با مقایسه نتایج سه مسئله‌ی رگرسیون، طبقه‌بندی و خوشه‌بندی، می‌توان دید که هر روش استخراج یا انتخاب ویژگی اثر متفاوتی بر مدل‌ها گذاشته است و برخی از آن‌ها در ایجاد ویژگی‌های مؤثرتر و مستقل‌تر موفق‌تر بوده‌اند.

۱ .مسئله رگرسیون Linear وSGDRegressor

استفاده از **PCA**منجر به کاهش میانگین و انحراف معیار ضرایب شد (Mean|coef|) از 0.417 به 0.212)، که نشان‌دهنده‌ی افزایش **پایداری ضرایب** و کاهش اثر همخطی میان ویژگی‌هاست. در SGDRegressor نیز منحنی خطا نشان داد که داده‌های تبدیل‌شده با PCA سریع‌تر به همگرایی می‌رسند، زیرا استقلال ویژگی‌ها جهت گرادیان را یکنواخت‌تر کرده است. در نتیجه، از منظر **تحلیل بهینه‌سازی**، PCAپایداری و سرعت همگرایی را بهبود داده است.

۲ .مسئله طبقه‌بندی KNN وRandomForest

- درKNN ، دقت در همه روش‌ها ثابت ماند (0.956) اما زمان پیش‌بینی پس از PCA تقریباً نصف شد (از 0.0020 به 0.001 ثانیه). این کاهش نشان می‌دهد که کاهش ابعاد باعث ساده‌تر شدن فضای فاصله‌ای و افزایش سرعت محاسبات شده است، هرچند تأثیری بر کیفیت تصمیم‌گیری نداشت. در مقابل، درRandomForest ، داده‌های اصلی (بدون حذف همخطی) بهترین عملکرد را داشتند(Accuracy=0.956) ، در حالی که PCA کمی دقت را کاهش داد (0.921). این نشان می‌دهد که مدل‌های درختی **به‌صورت ذاتی نسبت به همخطی مقاوم هستند** و نیازی به تبدیل فضای ویژگی ندارند.

۳ .مسئله خوشه‌بندی:

در این بخش، هر دو روش PCA و SelectKBest منجر به خوشه‌بندی بهتر شدند، اما **SelectKBest**بهترین نتایج را داشت:

- Inertiaاز 139 (Original) به 18 (SelectKBest) کاهش یافت.
- Silhouetteاز 0.46 به 0.67 افزایش پیدا کرد.

این اعداد نشان می‌دهد که SelectKBest توانسته است ویژگی‌های مؤثرتر برای جداسازی خوشه‌ها انتخاب کند و بنابراین از دیدگاه **کیفیت ویژگی‌ها** نسبت به PCA موفق‌تر بوده است.

نتیجه‌گیری:

- از دیدگاه **ایجاد ویژگی‌های مستقل‌تر، PCA**بهترین عملکرد را داشت، زیرا ساختار داده را به مؤلفه‌های غیرهم‌خطی تبدیل کرد و پایداری ضرایب و سرعت همگرایی را افزایش داد.
- از دیدگاه **ایجاد ویژگی‌های مؤثرتر برای جداسازی نمونه‌ها، SelectKBest**درKMeans و RandomForest نتایج قوی‌تری ارائه داد، چون ویژگی‌هایی را انتخاب می‌کند که بیشترین ارتباط آماری با برچسب هدف دارند.
- **ادغام استخراج ویژگی) مثل (PCA با تحلیل بهینه‌سازی (مانند رفتار ضرایب و همگرایی)** برای یک متخصص داده‌کاوی اهمیت دارد، زیرا درک تعامل بین ساختار داده و رفتار الگوریتم‌های یادگیری را ممکن می‌سازد. به‌طور خاص، چنین ادغامی کمک می‌کند تا تصمیم‌گیری بهینه بین «ساده‌سازی داده» و «حفظ اطلاعات مؤثر» انجام شود، که در کاربردهای واقعی تفاوت چشمگیری در دقت، پایداری و سرعت یادگیری ایجاد می‌کند.

منابع و مراجع

[1] منابع اصلی درس داده کاوی

[2] دانش خودم از درس های ماشبه که پاس کردم با همین مطالب این درس

Introduction to Machine Learning with Python – Andreas Müller & Sarah Guido

Data Mining: Concepts and Techniques – Han, Kamber, Pei

[6] Pattern Recognition and Machine Learning – Christopher M. Bishop

Abstract

This page is accurate translation from Persian abstract into English.

Key Words: Write a *3* to *5* KeyWords is essential.

Amirkabir University of Technology
(Tehran Polytechnic)

… Department …

MSc or PhD Thesis

Title of Thesis

By
Name

Supervisor
Dr.

Advisor
Dr.

Month & Year