به نام خدا

تمرین سری دوم

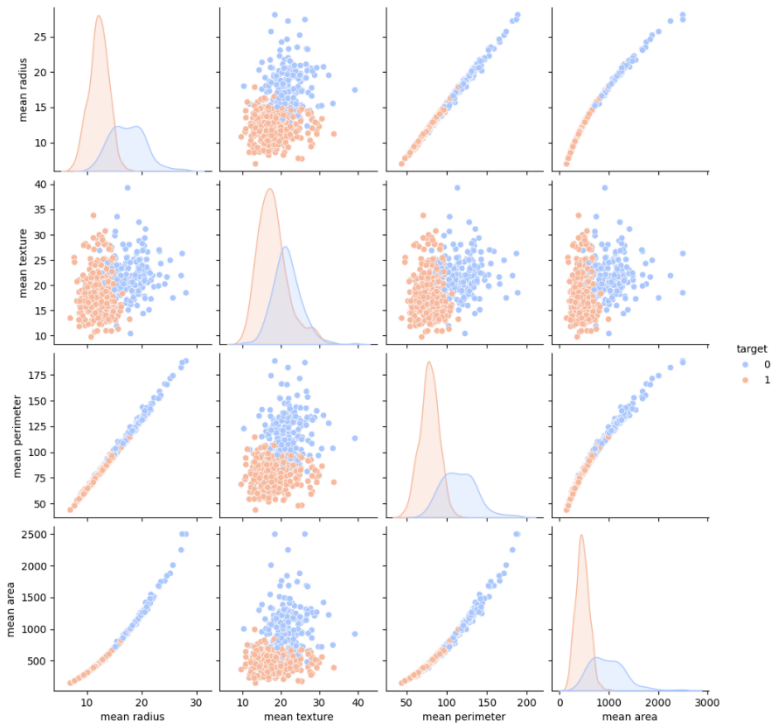
ائلدار صمدزاده طریقت ۴۰۰۳۳۰۴۴

### ۲-۲- بر اساس خروجی Pairplot

۱. کدام ویژگی‌ها بیشترین جداسازی را بین کلاس‌ها نشان می‌دهند؟ چرا این ویژگی‌ها می‌توانند برای دسته‌بندی مهم باشند؟
۲. آیا بین دو ویژگی خاص همبستگی قوی مشاهده می‌شود؟ این همبستگی چگونه می‌تواند بر عملکرد مدل تأثیر بگذارد؟
۳. آیا ویژگی‌های انتخاب‌شده توزیع یکسانی دارند، یا برخی از آن‌ها دارای مقادیر پرت هستند؟ این مسئله چگونه بر عملکرد مدل تأثیر می‌گذارد؟

۱. با توجه به pairplot ، mean radius و mean perimeter بیشترین همبستگی خطی را بین دیگر فیچرها دارند لذا نمودارشان هم به خط درجه یک نزدیکتر و همچنین نقاط در نزدیکی یکدیگر قرار دارند و پخش نیستند. در درجه دوم radius vs area و perimeter vs area همبستگی قوی‌ای دارند. دلیل اهمیت این ویژگی‌ها، قابلیت طبقه بندی ساده توسط انهاست. مثلا به جای عکس پزشکی، از این پارامترها راحت میتوان نوع تومور را حدس زد. از طرفی همبستگی بالای انها با هدف، انها را به شدت برای پیشبینی مناسب می‌کند.

۲. ویژگی‌هایی که به شدت با یکدیگر همبستگی دارند، معمولاً حاوی اطلاعات ارزشمندی برای وظایف طبقه‌بندی هستند، اما نیاز به توجه دقیق دارند. این همبستگی‌ها می‌توانند باعث ایجاد وابستگی در ویژگی‌ها شوند، به طوری که اطلاعات مشترک بین آن‌ها منجر به افزایش توان پیش‌بینی شود. اما اگر همبستگی بیش از حد باشد، می‌تواند منجر به افزونگی اطلاعات شود، جایی که ویژگی‌ها بینش‌های منحصربه‌فردی به مدل ارائه نمی‌دهند. علاوه بر این، ویژگی‌های به شدت مرتبط می‌توانند موجب بیش برازش مدل شوند، به این معنی که مدل ممکن است تأثیر زیادی از این ویژگی‌ها بگیرد و در داده‌های جدید عملکرد خوبی نداشته باشد. در این موارد، استفاده از تکنیک‌هایی مانند کاهش ابعاد (مانند تحلیل مولفه‌های اصلی) یا حذف یک ویژگی در مرحله پیش‌پردازش می‌تواند به کاهش افزونگی و ساده‌سازی مدل کمک کند.

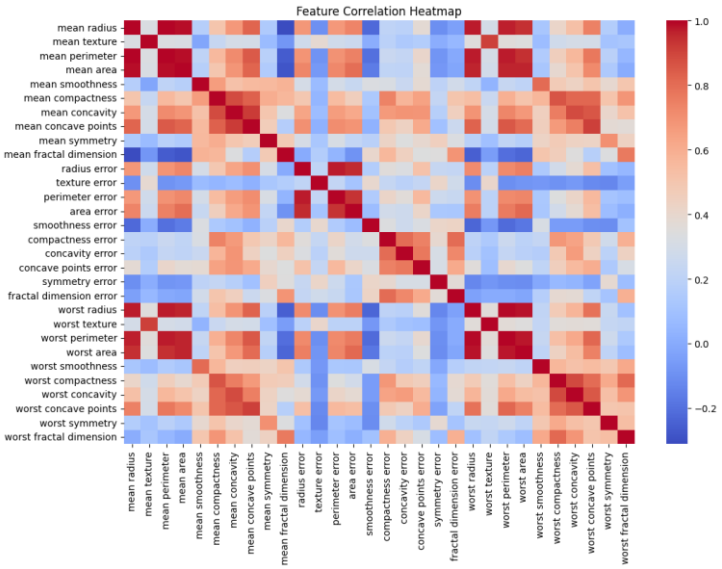



۳. برخی ویژگی‌های نمایش داده شده در نمودار دارای مقادیر نامتعارف یا توزیع غیر یکنواخت هستند. به طور خاص، ویژگی‌هایی مانند area ممکن است پخش وسیع‌تر و مقادیر غیرعادی بیشتری داشته باشند. این موارد می‌توانند فرآیند تحلیل و مدل‌سازی را تحت تأثیر قرار دهند. مقادیر نامتعارف می‌توانند بر عملکرد مدل تأثیر منفی بگذارند، به ویژه اگر الگوریتم انتخاب شده به این مقادیر حساس باشد، مانند رگرسیون خطی یا الگوریتم K-nearest neighbors. همچنین، توزیع غیر یکنواخت ویژگی‌ها می‌تواند روابط بین متغیرها را به اشتباه نشان دهد یا توانایی مدل در تعمیم الگوها را محدود کند. با استفاده از روش‌های پیش‌پردازشی مانند نرمال‌سازی یا حذف مقادیر نامتعارف، می‌توان عملکرد مدل را بهبود بخشید.

### ۲-۳- بر اساس خروجی نقشه حرارتی همبستگی

۱. کدام ویژگی‌ها بیشترین همبستگی را با یکدیگر دارند و چگونه این موضوع می‌تواند بر انتخاب ویژگی‌های مدل تأثیر بگذارد؟
۲. آیا ویژگی‌هایی با همبستگی بسیار پایین یا نزدیک به صفر مشاهده می‌شود؟ این مسئله چه اطلاعاتی درباره استقلال ویژگی‌ها ارائه می‌دهد؟
۳. اگر دو یا چند ویژگی همبستگی بسیار بالایی داشته باشند، چگونه می‌توان این مسئله را در فرآیند پیش‌پردازش داده مدیریت کرد؟

۱. بر اساس نقشه گرمایی ارائه شده، بالاترین میزان همبستگی مثبت بین ویژگی‌های *شعاع متوسط* و *محیط متوسط* مشاهده می‌شود، همچنین بین *مساحت متوسط* و *محیط متوسط* نیز همبستگی بالایی وجود دارد. این روابط نشان می‌دهند که این ویژگی‌ها ممکن است اطلاعات مشترکی داشته باشند و به طور مشابه به پیش‌بینی‌های مدل کمک کنند.




۲. ویژگی‌هایی مانند *اشتباه بافت* و *نرمی متوسط* دارای همبستگی نزدیک به صفر یا بسیار کم هستند. این نشان‌دهنده استقلال آنها است که می‌تواند در ایجاد تنوع در ورودی مدل پیش‌بینی مفید باشد.

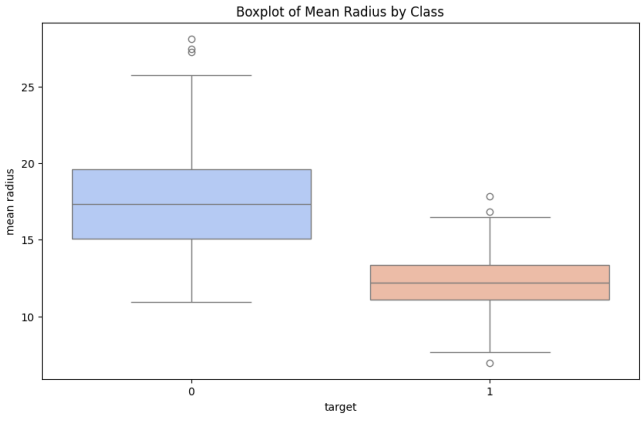
۳. برای مدیریت ویژگی‌های با همبستگی بالا در مرحله پیش‌پردازش، می‌توان ویژگی‌هایی مانند *مساحت متوسط* و *محیط متوسط* را حذف کرد تا افزونگی کاهش یابد. همچنین می‌توان از تکنیک‌های کاهش ابعاد مانند تحلیل مؤلفه‌های اصلی (PCA) برای ترکیب آنها به یک ویژگی نماینده استفاده کرد. علاوه بر این، روش‌های منظم‌سازی (مانند رگرسیون لاسو) می‌توانند کمک کنند تا مدل به طور مؤثر ویژگی‌ها را انتخاب کند و همبستگی‌ها را مدیریت نماید.

---

### ۲-۴- بر اساس خروجی Boxplot

۱. میانه و دامنه بین‌چهارکی (IQR) ویژگی "Mean Radius" در دو کلاس چگونه مقایسه می‌شود و چه برداشتی می‌توان از آن داشت؟
۲. آیا نقاط پرت (outliers) در یکی از کلاس‌ها مشاهده می‌شود؟ این نقاط چگونه می‌توانند بر عملکرد مدل یادگیری ماشین تأثیر بگذارند؟
۳. آیا پراکندگی مقدار "Mean Radius" در کلاس‌های مختلف مشابه است یا تفاوت قابل‌توجهی دارد؟ این موضوع چه تأثیری بر تفکیک‌پذیری کلاس‌ها دارد؟

۱. بر اساس تحلیل باکس‌پلات ارائه شده، میانه و بازه بین چارکی (IQR) ویژگی "شعاع میانگین" را می‌توان بین دو کلاس مقایسه کرد تا روند مرکزی و پراکندگی این ویژگی را مشاهده کنیم. اگر یکی از کلاس‌ها دارای میانه بالاتر یا بازه بین چارکی گسترده‌تری باشد، این نشان می‌دهد که ارزش‌های مرکزی و پراکندگی بین گروه‌ها متفاوت است و این موضوع می‌تواند قابلیت تفکیک کلاس‌ها را افزایش دهد.



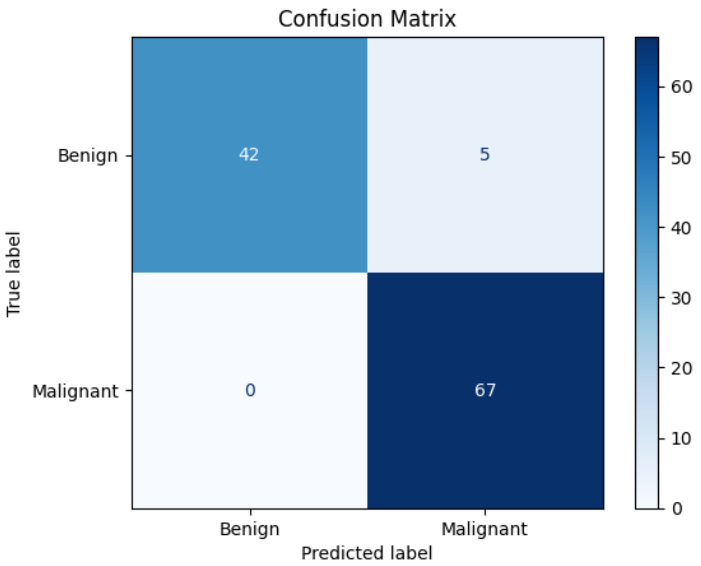

۲. اگر نقاط خارج از محدوده در یکی از کلاس‌ها وجود داشته باشند، ممکن است در آموزش مدل بایاس ایجاد کنند، به خصوص برای الگوریتم‌هایی که به توزیع داده حساس هستند. پردازش نقاط خارج از محدوده (مانند حذف یا تبدیل آنها) ممکن است برای بهبود عملکرد مدل ضروری باشد.

۳. مشاهده پراکندگی ویژگی "شعاع میانگین" در بین کلاس‌ها می‌تواند نشان دهد که آیا گسترش مقادیر این ویژگی مشابه است یا به طور قابل توجهی متفاوت. اگر این پراکندگی به شدت متفاوت باشد، ممکن است تمایز کلاس‌ها آسان‌تر شود، زیرا ویژگی "شعاع میانگین" مرزهای واضح‌تری برای کلاس‌ها ارائه می‌دهد.

---

**۵-۲- بر اساس خروجی کد KNN و ماتریس سردرگمی**

۱. چگونه تأثیر استانداردسازی داده‌ها را بر عملکرد مدل KNN ارزیابی می‌کنید؟ آیا مدل بدون استانداردسازی مدل عملکرد مشابهی خواهد داشت؟
۲. با توجه به ماتریس سردرگمی، مدل KNN در تشخیص کدام کلاس (خوش‌خیم یا بدخیم) عملکرد بهتری دارد؟ این موضوع چه تأثیری بر کاربرد مدل در تشخیص پزشکی دارد؟
۳. چگونه تغییر تعداد همسایه‌ها (n_neighbors) در مدل KNN می‌تواند بر دقت و نرخ خطای مدل تأثیر بگذارد؟ آیا مقدار بهینه‌ای برای آن وجود دارد؟




۱. استانداردسازی داده‌ها باعث یکسانسازی تاثیر ویژگی‌ها در اموزش مدل میشود. درواقع توزیع هر ویژگی، میانگین، مین و ماکس و دیگر خصوصیات انها فرق دارد لذا انها را استاندارد میکنیم که تاثیرشان یکسان شود. مدل بدون این حالت ارزش بیشتری برای ویژگی‌های با اندازه بیشتر قائل می‌شود که ممکن است اصلا خوب نباشه.
۲. در تشخیص بدخیم بهتر عملکرده چون که falsepositiveهایمان صفراند. این موضوع خیلی خوب است چون پیش‌بینی truenegative تاثیر مخرب کمتری نسبت به falsepositive دارد. اگر بیمار بدخیم را خوشخیم تشخیص داده و پیشگیری‌های لازم صورت نگیرد، جان مریض در خطر خواهد بود و این تماما به دلیل تشخیص اشتباه مدل بوده.
۳. افزایش مقدار n_neighbors معمولاً باعث می‌شود مدل پایدارتر شود زیرا برای طبقه‌بندی تعداد بیشتری از نقاط داده را در نظر می‌گیرد. با این حال، این امر ممکن است تمایزها را کاهش دهد و دقت را به ویژه در کلاس‌هایی با مرزهای ظریف کاهش دهد. از طرف دیگر، کاهش مقدار n_neighbors تصمیمات را دقیق‌تر می‌کند زیرا همسایه‌های کمتری را در نظر می‌گیرد، اما ممکن است مدل را نسبت به نویز حساس کند و منجر به بیش‌برازش داده‌ها شود. تعیین مقدار بهینه برای n_neighbors معمولاً شامل تنظیم این پارامتر با استفاده از تکنیک‌هایی مانند اعتبارسنجی متقابل (cross-validation) است. این فرآیند به تعادل دقت و نرخ خطا کمک می‌کند و بهترین مقدار را برای مجموعه داده‌های خاص شناسایی می‌کند.

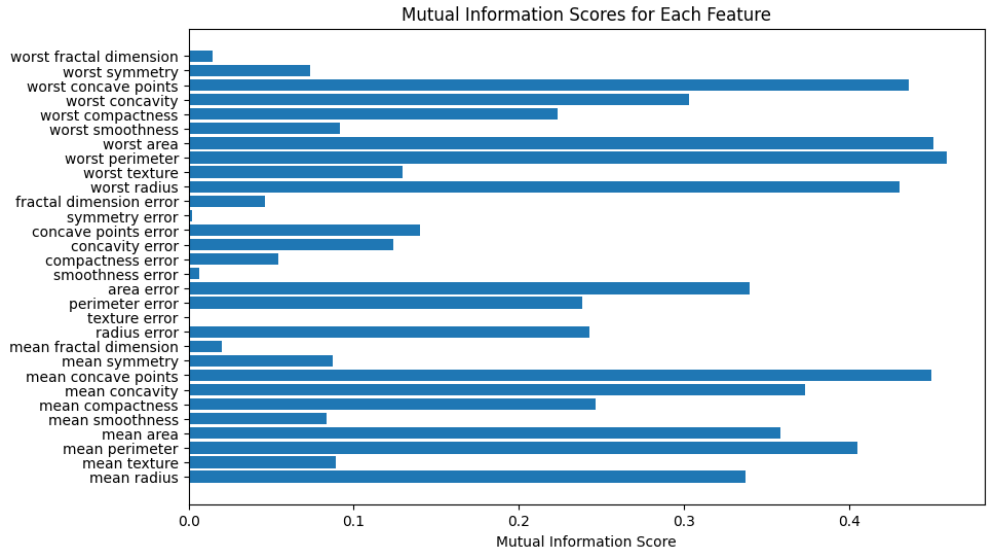
---

**۶-۲- بر اساس انتخاب ویژگی‌ها (Feature Selection) با استفاده از روش‌های مختلف**

۱. در روش انتخاب ویژگی با استفاده از "Mutual Information" ، چه ویژگی‌هایی به عنوان مهم‌ترین ویژگی‌ها انتخاب می‌شوند و چرا این ویژگی‌ها می‌توانند برای مدل مفیدتر باشند؟
۲. در روش "Recursive Feature Elimination" (RFE) ، چگونه تعداد ویژگی‌های انتخاب‌شده و مدل پایه (در اینجا رگرسیون لجستیک) بر عملکرد نهایی مدل تأثیر می‌گذارد؟ اگر تعداد ویژگی‌ها را افزایش دهیم، چه تغییراتی در دقت مدل مشاهده می‌شود؟
۳. در روش "SelectKBest" با استفاده از تست کای‌اسکوئر (Chi-square) ، چرا این روش به‌ویژه برای انتخاب ویژگی‌های گسسته مناسب است؟ چطور می‌توان این روش را برای ویژگی‌های پیوسته تطبیق داد؟

۱. در روش Mi طبق فرمول، امتیاز همه فیچرها محاسبه شده و سپس ۱۰ ویژگی با امتیاز بالاتر انتخاب می‌شوند. در اینجا به ترتیب: 'worst concave points' 'worst radius' 'mean perimeter' 'mean ' 'perimeter' 'worst area' 'mean concave points' 'mean area' 'mean radius' 'area error' 'concavity'.
۲. تعداد ویژگی‌های انتخابی در RFE تعیین‌کننده است. مقدار n_features_to_select=10 مشخص می‌کند که در نهایت ۱۰ ویژگی برتر انتخاب شوند. اگر تعداد کمی از ویژگی‌ها انتخاب شوند، ممکن است مدل اطلاعات مهمی را از دست بدهد و باعث کاهش دقت شود. در مقابل، انتخاب تعداد زیادی ویژگی می‌تواند نویز یا اطلاعات زائد را وارد کند و عملکرد مدل را تحت تأثیر قرار دهد و شاید مدل دچار بیش برازش شود. روش RFE به مدل پایه

نیز وابسته است. در اینجا از رگرسیون لجستیک به عنوان مدل پایه استفاده شده است که اهمیت ویژگی‌ها را بر اساس روابط خطی بین ویژگی‌ها و متغیر هدف را نشان می‌دهد. اگر از مدل دیگری مانند جنگل تصادفی استفاده شود، رتبه‌بندی ویژگی‌ها متفاوت خواهد بود زیرا این مدل‌ها روابط غیرخطی و تعاملات بین ویژگی‌ها را در نظر می‌گیرند.




۳. کد SelectKBest با استفاده از آزمون کای به توان دو برای انتخاب ویژگی‌های مرتبط‌تر با متغیر هدف مناسب است. آزمون کای به توان دو رابطه بین ویژگی‌های دسته‌ای و متغیر هدف را ارزیابی می‌کند. زیرا فرض می‌کند که ویژگی‌ها دسته‌ای هستند و ارتباط آن‌ها با متغیر هدف را بررسی می‌کند. برای ویژگی‌های پیوسته، آزمون کای دو مستقیماً قابل اعمال نیست، زیرا به داده‌های دسته‌ای نیاز دارد. لذا، می‌توان ابتدا آن‌ها را دسته‌بندی کرد، مثلاً با تقسیم به بین‌ها. این فرآیند داده‌های پیوسته را به داده‌های دسته‌ای تبدیل می‌کند و امکان استفاده از آزمون کای دو را فراهم می‌کند.

---

**۷-۲- در روش "SelectKBest" با استفاده از تست کای‌اسکوئر(Chi-square) ، خطایی ظاهر می شود، علت این خطا چیست؟**

۱. چرا روش) "Chi-square" کای‌اسکوئر) برای انتخاب ویژگی‌ها نیاز به داده‌های غیرمنفی دارد و چرا استانداردسازی داده‌ها ممکن است این الزامات را نقض کند؟
۲. کد را اصلاح کرده و مجددا اجرا کنید.
۳. چه تفاوت‌هایی بین روش‌های انتخاب ویژگی مانند "Chi-square" و "ANOVA F-value" وجود دارد و کدام‌یک برای داده‌های پیوسته و گسسته مناسب‌تر است؟ چرا باید هنگام انتخاب روش‌ها ویژگی‌های داده‌ها را در نظر بگیریم؟

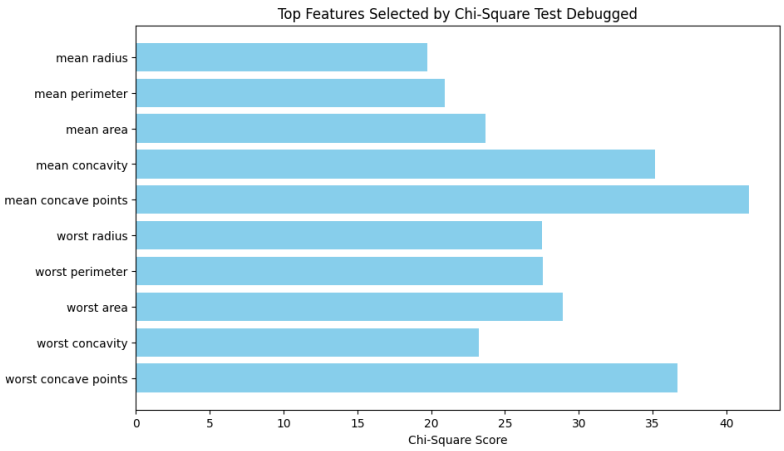
۱. آزمون کای اسکوئر آماری را بر اساس فرکانس های مشاهده شده و مورد انتظار در داده های گسسته محاسبه می کند. از آنجایی که مقادیر منفی فرکانس های معتبر را نشان نمی دهند، این روش برای محاسبه معنی دار به داده های غیرمنفی نیاز دارد. اگر ویژگی ها به دلیل مراحل پیش پردازش مانند استانداردسازی شامل مقادیر منفی باشند، این الزام نقض می شود. استانداردسازی می تواند با تغییر میانگین ویژگی ها به صفر، مقادیر منفی ایجاد کند و Chi-Square را بدون تبدیل داده ها به مقیاس های غیر منفی نامناسب کند.

۲.

```
# Apply Chi-Square feature selection
```

```
k = 10  # Number of top features to select
chi_selector = SelectKBest(score_func=chi2, k=k)  # Initialize SelectKBest with Chi-Square test
X_train_chi2 = chi_selector.fit_transform(X_train_scaled, y_train)  # Fit and transform the training data
X_test_chi2 = chi_selector.transform(X_test_scaled)  # Apply the same transformation to the test data
# Extract feature names and scores
selected_features = X.columns[chi_selector.get_support()]  # Get names of selected features
chi_scores = chi_selector.scores_  # Chi-Squared scores for all features
# Create a DataFrame for better visualization
chi2_results = pd.DataFrame({
    'Feature': X.columns,
    'Chi-Square Score': chi_scores
}).sort_values(by='Chi-Square Score', ascending=False)
```

۳. آزمون کای اسکوئر: ارتباط بین ویژگی های طبقه ای و متغیر هدف را ارزیابی می کند. ایده آل برای داده های گسسته و غیر منفی. ANOVA F-Test: میانگین بین گروه ها را برای ارزیابی واریانس مقایسه می کند و آن را برای داده های پیوسته مناسب می کند. چرا انواع داده ها را در نظر بگیرید: ویژگی های ویژگی ها (به عنوان مثال، گسسته یا پیوسته) تعیین می کند که کدام آزمون مناسب تر است. انتخاب یک روش نامناسب خطر تحریف نتایج یا ایجاد سوگیری را به همراه دارد.




---

**۲-۸- بر اساس نتایج حاصل از سه روش انتخاب ویژگی Mutual Information, Recursive Feature Elimination (RFE), و: SelectKBest (ANOVA F-value))**

۱. با توجه به نتایج حاصل از سه روش انتخاب ویژگی مختلف Mutual Information ، RFE و SelectKBest ، ویژگی‌هایی که به‌طور مشترک توسط این سه روش انتخاب شده‌اند، چه ویژگی‌هایی هستند؟ چرا این ویژگی‌ها احتمالاً برای پیش‌بینی بهتر مدل اهمیت دارند؟

۲. پس از اعمال هر یک از روش‌های انتخاب ویژگی، چگونه دقت مدل KNN یا هر مدل دیگری که انتخاب می‌کنید تغییر می‌کند؟ آیا انتخاب ویژگی‌ها باعث بهبود عملکرد مدل می‌شود یا خیر؟ دلیل این تغییرات چیست؟

۳. آیا می‌توانید توضیح دهید که چرا برخی ویژگی‌ها توسط یک روش انتخاب ویژگی انتخاب می‌شوند اما توسط روش‌های دیگر انتخاب نمی‌شوند؟ برای مثال، چرا برخی ویژگی‌ها ممکن است توسط **Mutual Information** انتخاب شوند، در حالی که **RFE** یا **SelectKBest** آن‌ها را کنار می‌گذارند؟ این تفاوت‌ها ممکن است نشانه‌ای از کدام ویژگی‌ها در داده‌ها باشد؟

۱. با توجه به نتایجی که در تصاویر پایین اورده شده، فیچرهای 0,2,7,20,22,23,27 در هر سه روش مشاهده میشوند. اهمیت این فیچرها از ان جهت است که با توجه به هر سه روش انتخاب فیچر، درصد بالایی از همکاری در اموزش مدل و ارتباط بهتری با خروجی دارند. لذا تاثیر انها در اموزش و درستی اموزش مدل کلیدی است.

| | Feature | CHI Score |
|---|---|---|
| 7 | mean concave points | 41.549482 |
| 27 | worst concave points | 36.705723 |
| 6 | mean concavity | 35.144083 |
| 23 | worst area | 28.913309 |
| 22 | worst perimeter | 27.573503 |
| 20 | worst radius | 27.518902 |
| 3 | mean area | 23.688277 |
| 26 | worst concavity | 23.238057 |
| 2 | mean perimeter | 20.944919 |
| 0 | mean radius | 19.757335 |

| | Feature | RFE Score |
|---|---|---|
| 0 | mean radius | 1 |
| 2 | mean perimeter | 1 |
| 7 | mean concave points | 1 |
| 26 | worst concavity | 1 |
| 20 | worst radius | 1 |
| 21 | worst texture | 1 |
| 22 | worst perimeter | 1 |
| 23 | worst area | 1 |
| 27 | worst concave points | 1 |
| 24 | worst smoothness | 1 |

| | Feature | MI Score |
|---|---|---|
| 22 | worst perimeter | 0.454553 |
| 7 | mean concave points | 0.450261 |
| 23 | worst area | 0.450016 |
| 27 | worst concave points | 0.433290 |
| 20 | worst radius | 0.432270 |
| 2 | mean perimeter | 0.406164 |
| 6 | mean concavity | 0.371855 |
| 3 | mean area | 0.357890 |
| 13 | area error | 0.338842 |
| 0 | mean radius | 0.338038 |

۲. تغییر عملکرد مدل پس از انتخاب ویژگی: از روش‌های انتخاب ویژگی برای کاهش ابعاد استفاده کنید، سپس مدل‌هایی مانند KNN را بر روی مجموعه داده‌های کاهش‌یافته آموزش دهید. دقت، صحت یا نمره F1 مدل را قبل و بعد از انتخاب ویژگی مقایسه کنید. مجموعه داده کاهش‌یافته باید به طور ایده‌آل عملکرد مدل را با حذف ویژگی‌های نامربوط و کاهش نویز بهبود بخشد.

۳. Mi وابستگی های غیر خطی را ثبت می کند. RFE ویژگی ها را بر اساس سهم آنها در پیش بینی های مدل رتبه بندی می کند و به طور مکرر ویژگی های کم اهمیت را حذف می کند. SelectKBest (Chi-Squared) به تداعی ویژگی های طبقه بندی شده با متغیر هدف متکی است. تفاوت ها به این دلیل بوجود می آیند که داده ها، وابستگی ها و فرضیات بر نحوه ارزیابی ویژگی ها تأثیر می گذارند.

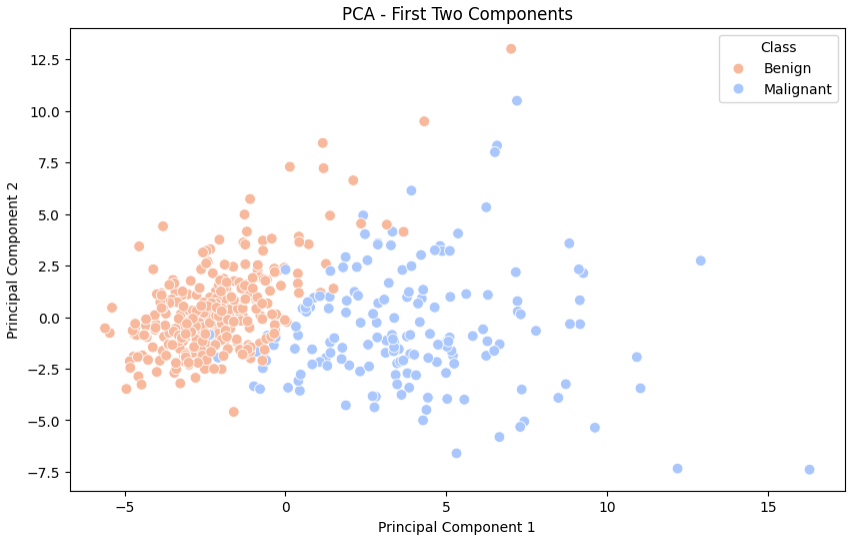
---

**۲-۹- کد مربوط به کاهش بعد PCA را اجرا کرده و خروجی نمایش داده شده را تشریح کنید. آیا برای انتخاب تعداد مؤلفه‌های اصلی در PCA معیار مشخصی وجود دارد؟**

تفسیر خروجی:

نمودار پراکندگی توزیع نقاط داده را بر حسب دو جزء اصلی اول نشان می دهد که بیشترین واریانس را در مجموعه داده به دست می آورند.

اگر کلاس ها (به عنوان مثال، "خوش خیم" و "بدخیم") خوشه های متمایز را تشکیل دهند، نشان می دهد که PCA با موفقیت ابعاد را کاهش داده است و در عین حال جداسازی کلاس را حفظ کرده است .




معیارهای انتخاب تعداد مؤلفه های اصلی:

یک روش متداول محاسبه نسبت واریانس توضیح داده شده با استفاده از pca.explained_variance_ratio_ است. این نشان می دهد که هر جزء اصلی چقدر واریانس دارد.

ترسیم واریانس تجمعی توضیح داده شده به تعیین تعداد بهینه مؤلفه ها کمک می کند. به عنوان مثال، اگر ۹۵ درصد از واریانس توسط ۱۰ مؤلفه اول گرفته شود، استفاده از n_components=10 منطقی است.

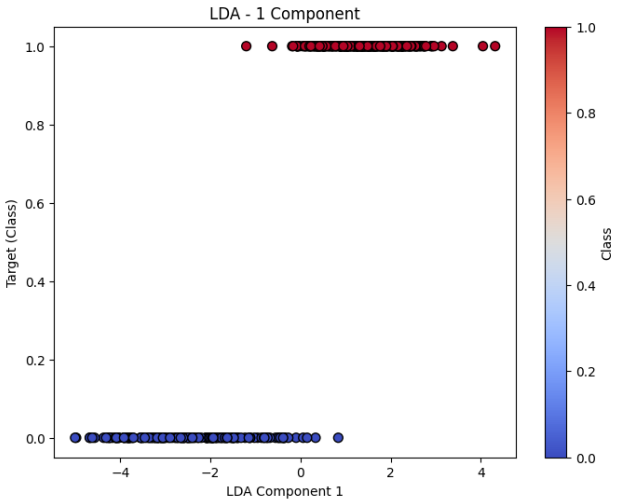
---

**۲-۱۰- چرا با اجرای کد مربوط به کاهش بعد با LDA خطا دریافت می شود؟ علت را تشریح کنید و کد را اصلاح کرده و خروجی را تشریح کنید.**

```
ValueError: n_components cannot be larger than min(n_features, n_classes - 1).
```

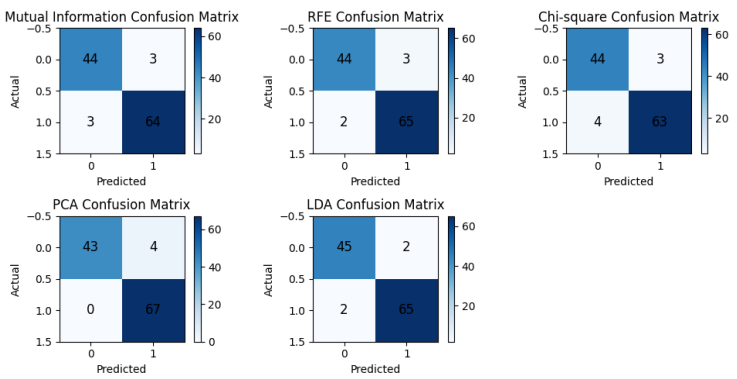
طبق ارور لاگ، تعداد کانپونننت‌های LDA نمیتواند از مینیموم فیچرها و یکی کمتر از تعداد کلاسها کمتر باشد و از انجایی که تعداد کلاس ها، ۲تاست، تعداد کلمپوننت‌ها میبایت حداکث ۱ باشد که در کد ۲ قرار داده شده بود.

با توجه به خروجی صلاح شده LDA، مشاهده میکنیم که کاهش ابعدا به گونه ای صورت گرفته که جداسازی کلاسها به راحتی انجام میگیرد. که نشان از درست عملکرد LDA ما دارد.




---

**۲-۱۱- با اجرای بلوک های اخر کد عملکرد KNN را با روشهای مختلف انتخاب ویژگی و کاهش بعد تشریح کنید.**




```
KNN Classifier Performance (Accuracy Scores): 0.9561
MI    : 0.9298
RFE   : 0.9737
F Test: 0.9386
PCA   : 0.9649
LDA   : 0.9649
```

با توجه به خروجی‌های مدلهای KNN طبق فیچر سلکشن و ریداکشن‌های مختلف، مدل در به غیر از حالت تست کا، در مابقی حالات دقت بالاتری نشان می‌دهد. همچنین طبق ماتریس‌های سردرگمی، در حالت اولیه که بالاتر اشاره شد، ۵ اشتباه در پیشبینی وجود داشت ولی به جز در حالت PCA و LDA، در حالات دیگر تعداد پردیکشن‌های اشتباه بیشتر شده. همچنین در falsepositive PCAهایمان همچنان صفراند که نشانه برتری pca بر lda در این مورد است.

**۱۲-۲- چه ایده هایی برای ترکیب استفاده از روش های انتخاب ویژگی و کاهش ابعاد به ذهنتان می رسد با جزئیات تشریح کنید و یک مورد را پیاده سازی و به انتهای کد اضافه کند و عملکرد KNN را با این روش ترکیبی جدید بسنجید.**

رویکرد: ترکیب RFE وPCA

- RFEبرای انتخاب ویژگی: از حذف ویژگی بازگشتی (RFE) برای کاهش مجموعه داده به مهم ترین ویژگی ها بر اساس ارتباط آنها با متغیر هدف استفاده کنید. این ویژگی های نامرتبطی را که ممکن است به عملکرد مدل کمک نکنند، کاهش می دهد.
- PCAبرای کاهش ابعاد PCA :را روی ویژگی های انتخاب شده اعمال کنید تا با طرح ریزی آنها در یک فضای با ابعاد پایین تر و در عین حال حفظ بیشتر واریانس، ابعاد را کاهش دهید.
- ارزیابی KNN: از مجموعه داده تبدیل شده برای آموزش و آزمایش مدل KNN استفاده کنید. عملکرد آن را با استفاده از دقت، دقت، یادآوری یا نمره F1 ارزیابی کنید.

```
#================================  QUESTION 12 !!!  ========================

pca = PCA(n_components=10)  # Reduce to 10 principal components
X_train_HYB = pca.fit_transform(X_train_rfe)
X_test_HYB = pca.transform(X_test_rfe)

acc_HYB, cm_HYB = evaluate_knn(X_train_HYB, X_test_HYB, y_train, y_test)

print("Accuracy of KNN with RFE + PCA:" , acc_HYB)
```

```
Accuracy of KNN with RFE + PCA: 0.9736842105263158
```

خروجی نهایی 0.973 است که به نسبت حالات قبلی پیشرفت داشته!

---

https://colab.research.google.com/drive/1tNuP65viYDH8AtiAv8mOgE3epC1FVCbJ?usp=sharing