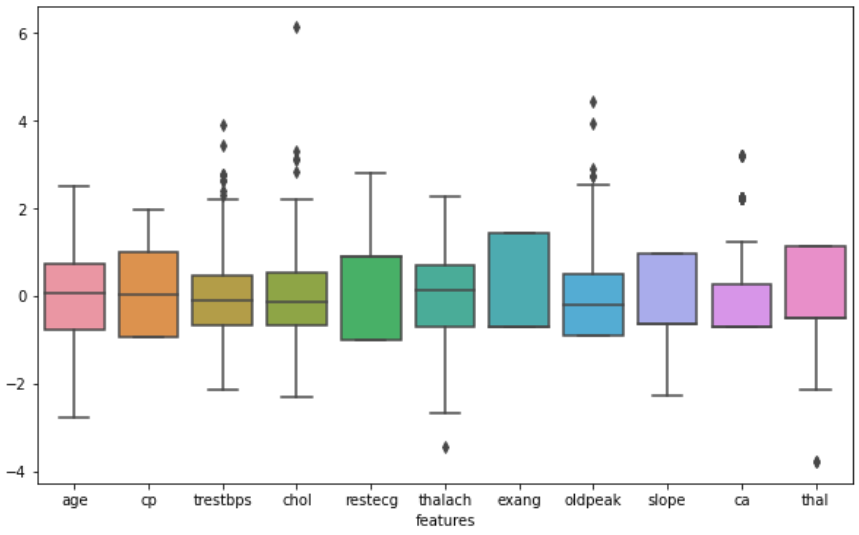
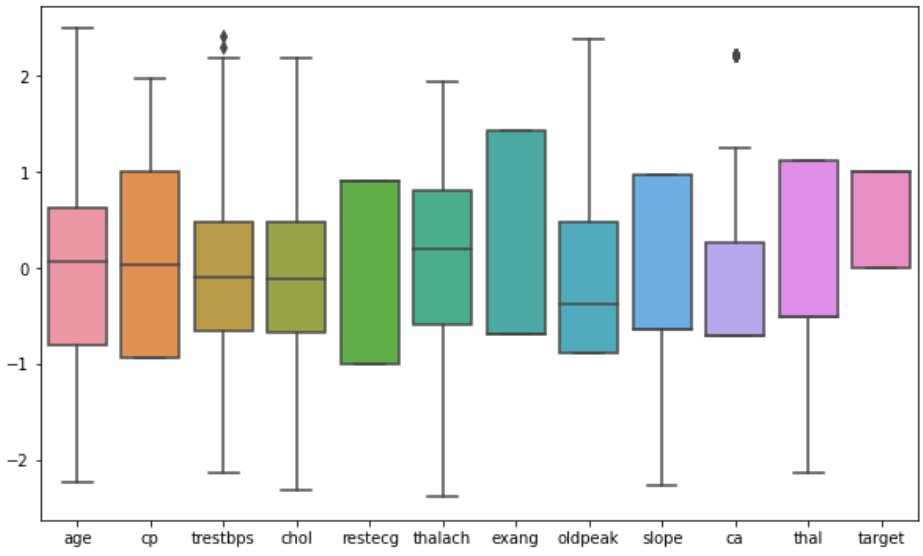
1. در ابتدا به بررسی دیتاست با استفاده از پکیج **pandas** بپردازید .

- **آیا داده پرت در دیتاست وجود دارد ؟ در صورت وجود آنها را حذف کنید .**

بله. شکل 1 نقاطی را نشان می‌دهد که فاصله بسیار زیادی از مرکز توزیع دارند که در واقع همان نقاط پرت موجود در داده‌ها می‌باشند. توجه داشته باشید که داده‌ها را قبلا استاندارد کرده ایم. برای حذف داده‌های پرت، هر نمونه‌ای که فاصله ای بیش از 2.5 از مبدا صفر داشت را از داده‌ها حذف کردیم. در شکل 2 مشاهده می‌کنید که پس از حذف این نقاط، تعداد نقاط پرت موجود در مجموعه داده‌ها کاهش یافته است.



شکل 1. نمودار جعبه ای داده ها قبل از حذف داده های پرت

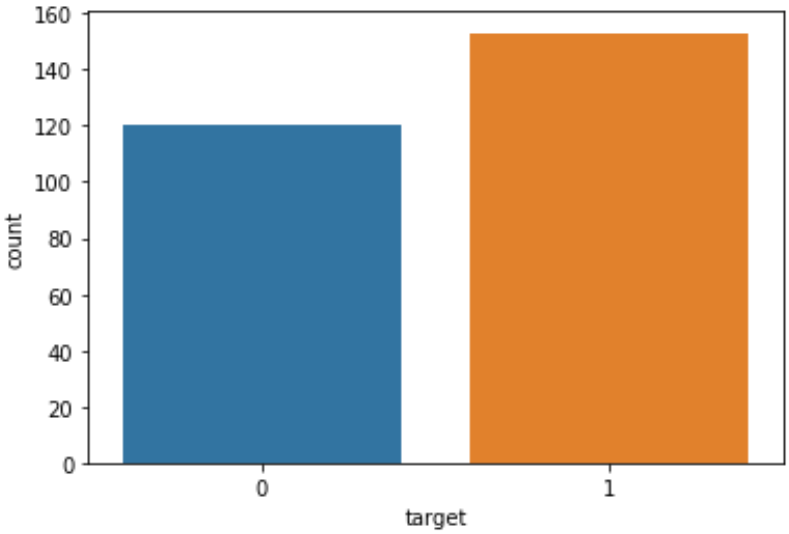


شکل 2. نمودار جعبه ای داده‌ها پس از حذف داده‌های پرت

- **بررسی کنید آیا تعداد نمونه ها در هر کلاس متوازن است ؟ (به صورت مختصر توضیحدهید اگر داده ها متوازن نباشد چه مشکلاتی ممکن است پیش بیاید وچه راه حلهایی برای آن وجود دارد).**

شکل 3 بیانگر تعداد رخداد هر یک از کلاس‌ها در کل داده‌ها می‌باشد. با مقایسه تعداد نمونه‌های موجود در دو کلاس موجود در می‌یابیم که داده‌ها متوازن هستند. چرا که تعداد رخداد کلاس‌ها تفاوت قابل توجهی با یکدیگر ندارند.

در مواردی که هزینه دسته‌بندی برای کلاس‌های مختلف با هم متفاوت باشد، داده نامتوازن مشکل-ساز می‌شود. از طرفب در این موارد، به دلیل نامتوازن بودن داده‌ها، نمی‌توان از معیار accuracy برای ارزیابی مدل بهره برد. سه معیار precision، recall و f1-score به عنوان جایگزینی مناسب برای حل این چالش مورد استفاده قرار می‌گیرند.



شکل 2. هیستوگرام مقادیر target

سوال 3. قضیه بیز را در حداقل یک پاراگراف بیان کنید .

- قضیه بیز: این قضیه را می‌توان به صورت زیر فرمول‌بندی کرد:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$P(A|B)$: احتمال رخداد A به شرط B

$P(A)$: احتمال رخداد A

$P(B)$: احتمال رخداد B

- سپس دسته بند های Gaussian Naive Bayes، Multinomial Naive Bayes، Bernoulli Naïve Bayes را با یکدیگر مقایسه کنید و بیان کنید هر کدام از این دسته بندها بیشتر در کجا کاربرد دارند.

Multinomial Naïve Bayes، برای پردازش داده‌هایی مناسب می‌باشد که در آن‌ها هر بردار ویژگی ، نشانگر مقادیر گسسته باشد. مثلا تعداد رخداد تصادف در هر روز.

Gaussian Naive Bayes زمانی مناسب است که داده‌ها دارای توزیع گاوسی باشند، و بیشتر برای پردازش داده‌های پیوسته مورد استفاده قرار می‌گیرد

Bernoulli Naïve Bayes: فقط برای تحلیل داده‌های باینری مناسب می‌باشد.

7. بررسی کنید که در سه معیار مطرح شده مدلی که با استفاده از پکیج ساخته اید و مدلی کهخود پیاده سازی کرده اید به چه صورتی عمل کرده اند.

	مدل پیاده‌سازی در سوال 5	مدل پیاده‌سازی شده در سوال 6
Precision	0.71	0.71
Recall	0.79	0.79
F1-score	0.74	0.74

جدول 1. مقادیر precision، recall و f1-score

همانطور که در جدول 1 می‌بینید، مدل پیاده‌سازی شده با استفاده از کتابخانه، تفاوت چندانی با مدل پیاده‌سازی شده از پایه ندارد.

8 . کلاسیفایر SVM را با استفاده از پکیج sklearn بر سه فیچر مطرح شده در سوال (۴) بااستفاده از داده های آموزشی ترین کنید . سپس بر روی داده های تست سه معیار Recall، Precision، F1 score را گزارش کنید .

```
precision = 0.70
recall = 0.75
f1-score = 0.75x
```

9. حداقل دو حالت مختلف را برای کرنل در SVM ساخته شده با پکیج در نظر بگیرید ونتایج آن را گزارش دهید . آیا کرنل های مختلف نتایج مختلفی ارایه دادند ؟ به صورتکلی علت استفاده از کرنل ها در SVM چیست ؟ توضیح دهید .

	کرنل خطی	کرنل rbf
Precision	0.76	0.63
Recall	0.79	0.75
F1-score	0.75	0.75

جدول 2. مقادیر precision، recall و f1-score

با توجه به جدول 2، کرنل خطی نتایج بهتری نسبت به rbf را بازگردانده است.

- کرنل‌ها به ما این امکان را می‌دهند که داده‌ها را با صرف هزینه کمتر و به شکلی کارآ به فضایی با ابعاد بالاتر ببریم.

- همانطور که اشاره شد، کاربرد کرنل افزایش ابعاد داده‌ها می‌باشد. در مواردی افزایش ابعاد و تغییر مختصات، باعث می‌شود تا داده‌ها قابلیت جداسازی بیشتری پیدا کنند. این مسئله برای الگوریتم svm که به دنبال بیشینه‌سازی فاصله بردارهای پشتیبان می‌باشد بسیار مهمی ست.

10 . دسته بند SVM را با استفاده از پکیج sklearn بسازید و با در نظر گرفتن کلیه فیچرهایدیتاست بر روی داده های آموزشی ترین کنید سپس نتایج را بر روی داده های تست ،ارزیابی کنید .

```
precision= 0.82
recall= 0.82
f1-score= 0.74
```

۱۱ . برای سوال (10)، یکبار مدل را با 5-fold Cross Validation اجرا کنید و نتایج راگزارش دهید .(در این جا برای فولد کردن داده ها از کل دیتاست استفاده میکنیم).

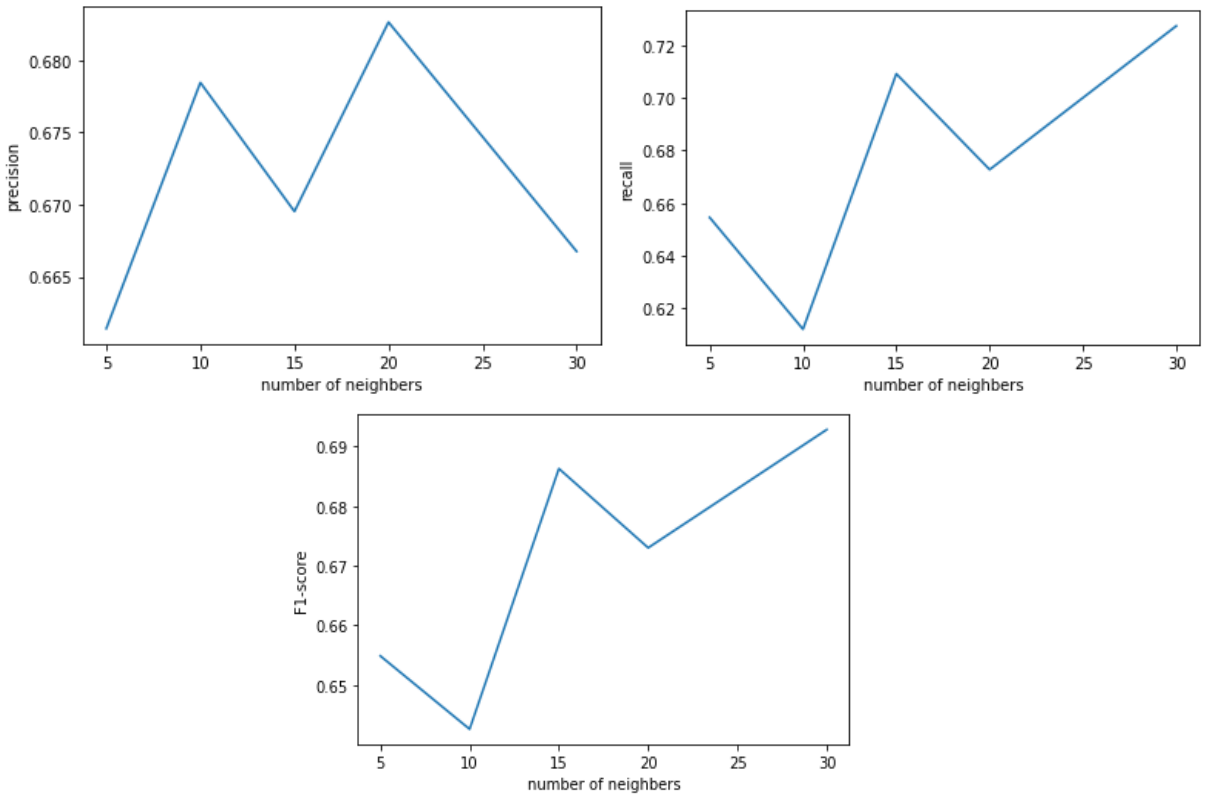
در این‌جا مقدار precision،recall و f1-score را برای هر فولد حساب کرده و میانگین هر کدام را در ادامه گزارش خواهیم کرد.

```
precision= 0.81
recall= 0.89
f1-score= 0.85
```

12 . با استفاده از پکیج sklearn دسته بند را K-NN [۳] را بسازید . با به کارگیری تمامی فیچرها موجود در دیتاست آموزشی ، مدل را ترین کنید سپس بر روی دیتاست تست ،ارزیابی کنید .

```
precision= 0.840234781031745
recall= 0.8545454545454545
f1-score= 0.845834275591738 2
```

13 . بررسی کنید در سوال (۱۲) تعداد همسایه ها k چه نقشی ایفا میکند ؟ زیاد شدن همسایه ها خوب است ؟ چگونه میتوان مشخص کرد چه تعداد همسایه برای مسئله ما مناسب است.



شکل 3. نمودار تغییر precision، recall و f1-score بر اساس تعداد همسایه‌های استفاده شده در knn-classidfier

همانطور که شکل 3 نشان می‌دهد، مقدار *precision* با افزایش تعداد همسایه‌ها تغییر معناداری نمی‌کند، ولی مقادیر *recall* و *f1-score* با افزایش تعداد همسایه‌ها افزایش می‌یابد.

برای مشخص کردن تعداد همسایه‌های بهینه، از روش *kfold-cross-validation* استفاده می‌شود. در این حالت تعدادی مناست می‌باشد که میانگین *precision* ، *recall* و *f1-score* مناسب را به ما ارایه دهد.

14. در سوال (۱۲) به جای استفاده از تمامی فیچرها فقط از سه فیچر trestbps، chol، thalach استفاده کنید و مدل را بسازید . سپس نتایج ارزیابی را گزارش دهید .

```
precision= 0.6667596060278986
```

```
recall= 0.7272727272727273
f1-score= 0.6928135974519773
```

15. تفاوت بین روش های کلاس بندی پارامتری و غیرپارامتری را به صورت خلاصه بیان کنید

. هر کدام بهتر است در چه مواقعی استفاده شوند ؟

در روش‌های پارامتریک، ما از قبل پیش فرض‌هایی را در مورد داده‌ها در نظر می‌گیریم. مثلا در Gaussian Naive Bayes پیش فرض ما این است که داده‌ها از توزیع گاوسی تبعیت می‌کنند. اما در روش‌های غیر پارامتریک از در نظر گرفتن چنین پیش‌فرض‌هایی اجتناب کرده و صرفا بر اساس شکل هندسی داده‌ها تحلیل را انجام می‌دهیم. از آن جا که در روش‌های غیرپارامتریک، تعداد پارامترها با افزایش ابعاد داده‌ها بشدت افزایش می‌یابد، این روش‌ها برای داده‌های با ابعاد بالا مناسب نیستند. ولی امتیاز روش‌های غیرپارامتریک این است که بهتر روی داده‌ها فیت می‌شوند، چون پیش فرض خاصی نسبت به ساختار داده‌ها ندارند.

16. (بخش امتیازی) معیار Matthews Correlation Coefficient(MCC) چیست ودر چه جاهایی استفاده میشود .

معیاری برای اندازه‌گیری کیفیت کلاسیفایرهای باینری می‌باشد. این معیار را می‌توان از فرمول زیر حساب کرد:

$$\mathbf{MCC} = \frac{\boldsymbol{TP} \times \boldsymbol{TN} - \boldsymbol{FP} \times \boldsymbol{FN}}{\sqrt{(\boldsymbol{TP} + \boldsymbol{FP})(\boldsymbol{TP} + \boldsymbol{FN})(\boldsymbol{TN} + \boldsymbol{FP})(\boldsymbol{TN} + \boldsymbol{FN})}}$$