گزارش سری تمرین 2

واحد درسی داده کاوی دوره الکترونیکی

جناب آقای دکتر خردپیشه

رحیم اکبری 99422028

00/01/21

فهرست مطالب

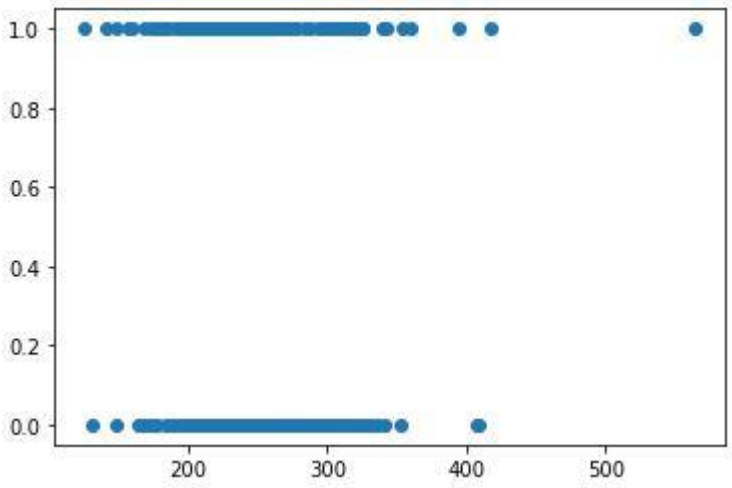
تشخیص بیماری قلبی

مجموعه داده‌ای که در اینجا داریم، مجموعه داده‌ای است متشکل از افرادی که دارای بیماری قلبی هستند یا نیستند. در شکل زیر نمایی از این مجموعه داده را می‌بینیم.

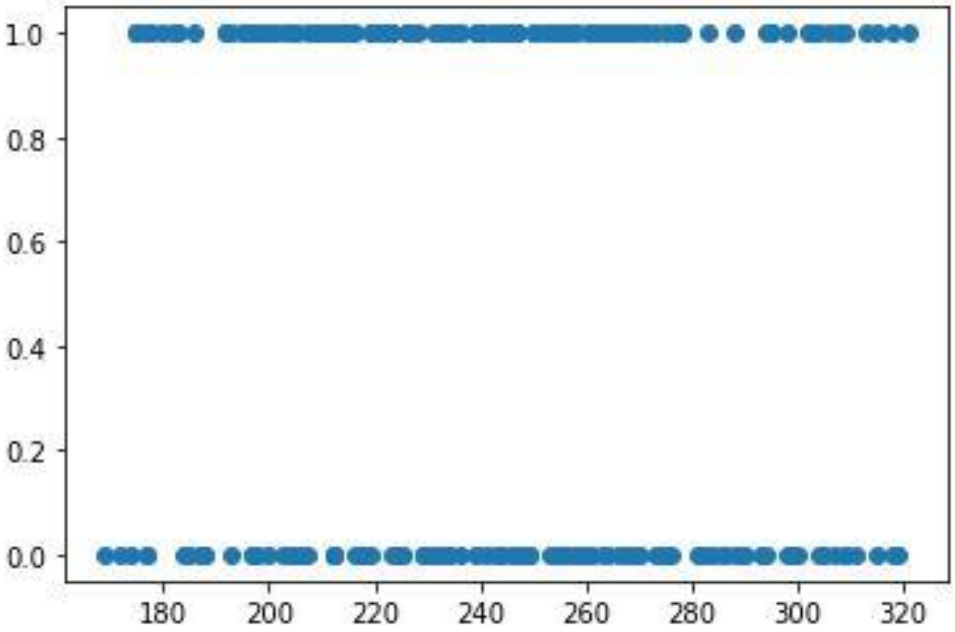
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

داده‌های پرت

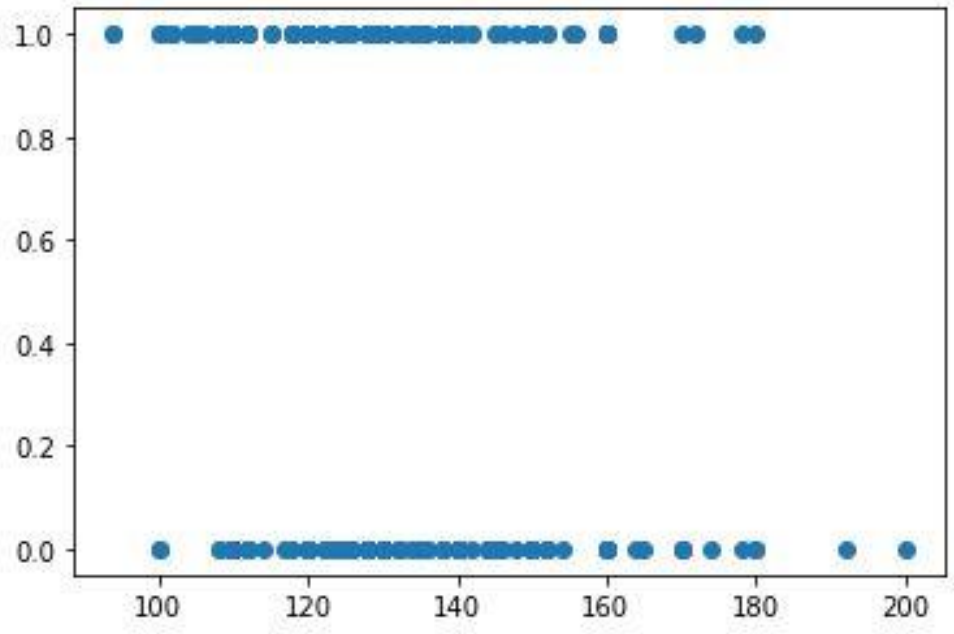
نمودار زیر نشان می‌دهد که در ستون chol داده پرت داریم. آن‌ها را حذف کرده و دوباره نمودار را بررسی می‌کنیم.



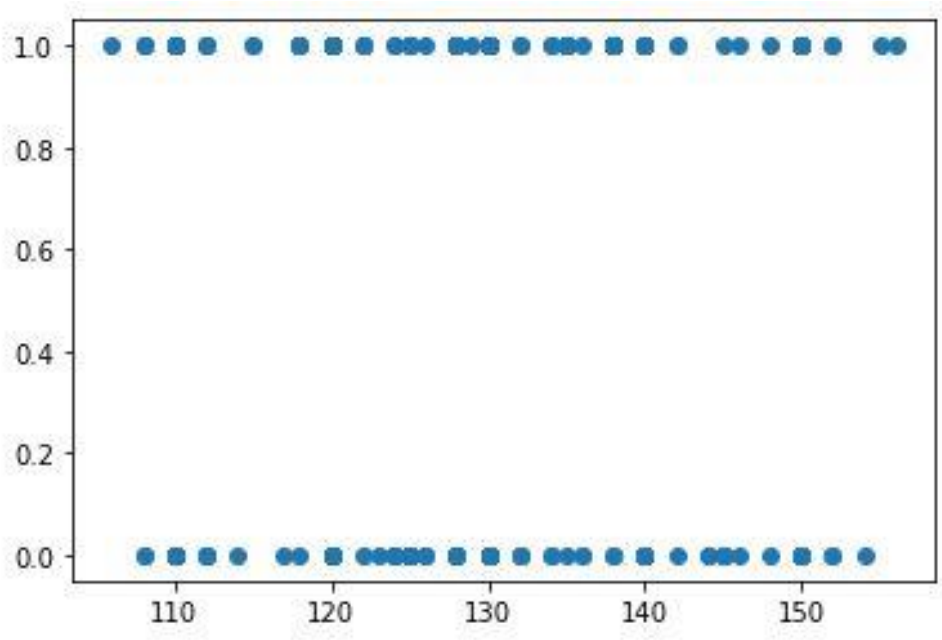
نمودار ستون chol پس از حذف داده‌های پرت بصورت زیر است.



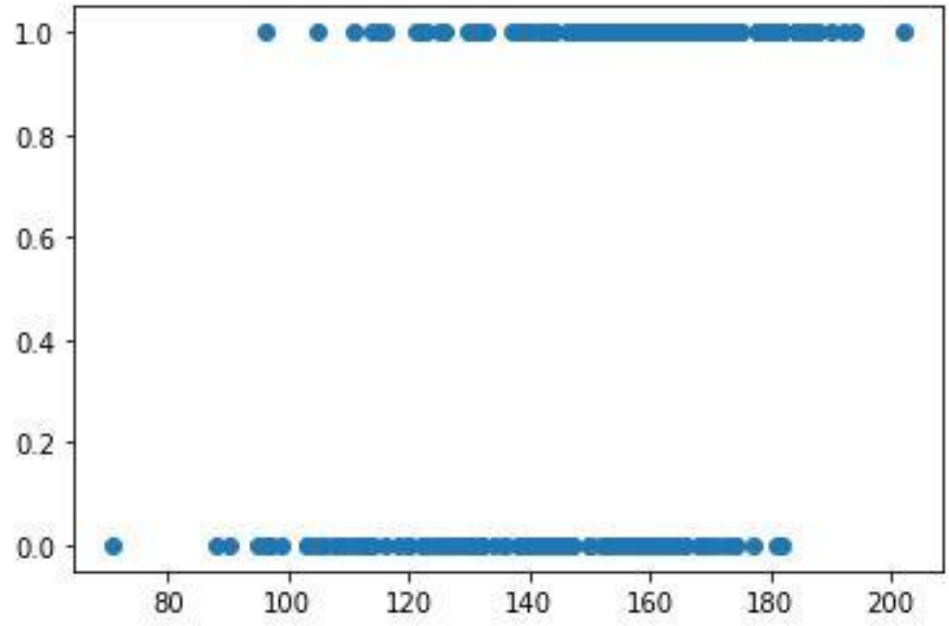
داده‌های پرت را در ستون trestbps بررسی می‌کنیم.



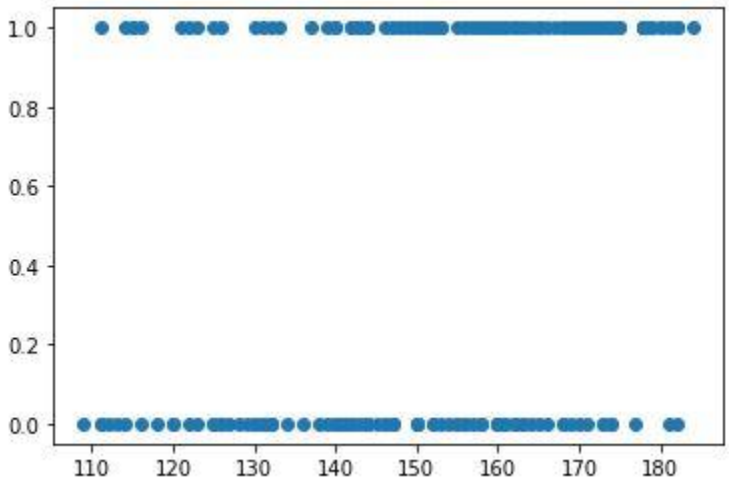
پس از حذف داده‌های پرت نمودار به شکل زیر می‌شود.



ستون thalach را برای داده‌های پرت بررسی می‌کنیم.



نمودار آن پس از حذف داده‌های پرت بصورت زیر است.



اطلاعات کلاس‌ها نشان می‌دهد که 151 نمونه در کلاس 1 و 124 نمونه در کلاس 0 داریم. داده‌ها متوازن است و مشکلی ندارد. اگر داده‌ها نامتوازن باشد آموزش مدل وابستگی بیشتری به کلاس بزرگتر پیدا می‌کند و هنگام تست ممکن است نمونه‌های کلاس کوچکتر را به اشتباه در کلاس دیگر دسته‌بندی کند. برای رفع این مشکل می‌توان از روی نمونه‌های کلاس بزرگتر با روش‌های مختلف به میزان کلاس کوچکتر نمونه برداشت تا متوازن شوند.

مدلسازی

ابتدا 80 درصد داده‌ها را به داده‌های آموزش و 20 درصد باقیمانده را به داده‌های تست اختصاص دادیم.

قضیه بیز

روشی برای دسته‌بندی پدیده‌ها، بر پایه احتمال وقوع یا عدم وقوع یک پدیده‌است و در نظریه احتمالات با اهمیت و پرکاربرد است. اگر برای فضای نمونه‌ای مفروضی بتوانیم چنان افرازی انتخاب کنیم که با دانستن اینکه کدامیک از پیشامدهای افراز شده رخ داده‌است، بخش مهمی از عدم قطعیت تقلیل می‌یابد.

این قضیه از آن جهت مفید است که می‌توان از طریق آن، احتمال یک پیشامد را با مشروط کردن نسبت به وقوع یا عدم وقوع یک پیشامد دیگر محاسبه کرد. در بسیاری از حالت‌ها، محاسبهٔ احتمال یک پیشامد به صورت مستقیم کاری دشوار است. با استفاده از این قضیه و مشروط کردن پیشامد مورد نظر نسبت به پیشامد دیگر، می‌توان احتمال مورد نظر را محاسبه کرد.

این رابطه به خاطر بزرگداشت توماس بیز فیلسوف انگلیسی به نام فرمول بیز معروف است.

اگر مشاهدات و داده‌ها از نوع پیوسته باشند، از مدل احتمالی با توزیع گاوسی یا نرمال برای متغیرهای مربوط به شواهد می‌توانید استفاده کنید. در این حالت هر دسته یا گروه دارای توزیع گاوسی است. به این ترتیب اگر k دسته یا کلاس داشته باشیم می‌توانیم برای هر دسته میانگین و واریانس را محاسبه کرده و پارامترهای توزیع نرمال را برای آن‌ها برآورد کنیم.

بیز ساده چندجمله‌ای، به عنوان یک دسته‌بند متنی بسیار به کار می‌آید. در این حالت برحسب مدل احتمالی یا توزیع چند جمله‌ای، برداری از n ویژگی برای یک مشاهده به صورت $X=(x1,x2,...,xn)$ با احتمالات $(p1,p2,...,pn)$ در نظر گرفته می‌شود. مشخص است که در این حالت بردار X بیانگر تعداد مشاهداتی است که ویژگی خاصی را دارا هستند.

به شکلی دسته¬بند نایو بیز برنولی بیشترین کاربرد را در دسته‌بندی متن‌های کوتاه داشته، به همین دلیل محبوبیت بیشتری نیز دارد. در این مدل در حالت چند متغیره، فرض بر این است که وجود یا ناموجود بودن یک ویژگی در نظر گرفته شود. برای مثال با توجه به یک لغتنامه مربوط به اصطلاحات ورزشی، متن دلخواهی مورد تجزیه و تحلیل قرار می‌گیرد و بررسی می‌شود که آیا کلمات مربوط به لغتنامه ورزشی در متن وجود دارند یا خیر.

مدل نایو بیز گاوسی را بدون استفاده از کتابخانه‌ها ساختیم. خروجی آن روی داده‌ها تست به شکل زیر است.

```
f1_score:gnb:  0.7719298245614035
precision:gnb:  0.7333333333333333
recall:gnb:  0.8148148148148148
```

سپس با استفاده از کتابخانه پایتون نایو بیز گاوسی را روی داده‌های آموزشی مدلسازی کردیم. خروجی روی داده‌های تست به شکل زیر است.

```
f1_score:gnb:  0.7719298245614035
precision:gnb:  0.7333333333333333
recall:gnb:  0.8148148148148148
```

همانطور که مشاهده می‌شود خروجی آن با مدلی که خودمان ساختیم تفاوتی ندارد.

سپس مدل svm را ساخته و بر روی داده‌های آموزشی آموزش دادیم. خروجی آن برای داده‌های تست به شکل زیر است.

```
f1_score:svm:  0.7868852459016393
precision:svm:  0.7058823529411765
recall:svm:  0.8888888888888888
```

مطابق با شکل بالا و شکل‌های قبلی، مدل svm در پارامترهای f1 و رکال بهتر و در پارامتر پرسیژن بدتر مدل نایو بیز گاوسی بوده است.

سپیس مدل svm را با کرنل زیگموید بر روی داده‌های آموزشی مدلسازی کردیم. خروجی بر روی داده‌های تست به شکل زیر است.

```
f1_score:svm:  0.6835443037974684
precision:svm:  0.5192307692307693
recall:svm:  1.0
```

سپس آن را با کرنل پولی آموزش دادیم. شکل زیر خروجی را نشان می‌دهد.

```
f1_score:svm:  0.7272727272727273
precision:svm:  0.7142857142857143
recall:svm:  0.7407407407407407
```

شکل‌ها نشان می‌دهند که با کرنل زیگوید پارامترهای f1 و پرسیژن کاهش و پارامتر رکال افزایش داشته است. با کرنل پولی نیز پارامترهای f1 و رکال کاهش و پارامتر پرسیژن افزایش داشته است.

از مدل 5-fold برای تقسیم‌بندی داده‌ها استفاده کرده و دوباره svm را بر روی داده‌های آموزش مدلسازی کردیم. خروجی به شکل زیر است.

```
f1_score:svm:  0.7142857142857143
precision:svm:  0.5681818181818182
recall:svm:  0.9615384615384616
```

پارامترهای f1 و پرسیژن کاهش و پارامتر رکال بهبود پیدا کرد.

مدل knn را برای k=3 آموزش دادیم. خروجی روی داده‌های تست به شکل زیر است.

```
f1_score:knn:  0.6440677966101696
precision:knn:  0.5757575757575758
recall:knn:  0.7307692307692307
```

تعداد همسایه‌ها را از 3 به 7 افزایش دادیم. خروجی به شکل زیر است.

```
f1_score:knn:  0.7419354838709676
precision:knn:  0.6388888888888888
recall:knn:  0.8846153846153846
```

با افزایش تعداد همسایه‌ها مدل عملکرد بهتری دارد، به این خاطر که هرچه تعداد همسایه‌ها بیشتر شود واریانس مدل بهتر شده و تاثیر داده‌های پرت و کم‌اثر در مدلسازی کمتر می‌شود.

مدل knn را بر روی همه ویژگی‌های chol، trestbps و thalach با تعداد همسایه 3 دوباره مدلسازی کردیم. خروجی به شکل زیر است.

```
f1_score:knn:  0.6666666666666667
precision:knn:  0.6129032258064516
recall:knn:  0.7307692307692307
```

نتیجه بالا نشان می‌دهد پارامترهای ارزیابی مدل افزایش پیدا کرد و این یعنی مدل عملکرد بهتری داشته است. در واقع این سه ویژگی تاثیر مهمتری نسبت به سایر ویژگی‌ها داشته‌اند.

مدل پارامتری و غیرپارامتری

یک مدل یادگیری که داده‌ها را با مجموعه ای از پارامترهای با اندازه ثابت (مستقل از تعداد نمونه های آموزشی) خلاصه می‌کند، یک مدل پارامتری نامیده می شود. مهم نیست که چه مقدار داده ای که در یک مدل پارامتری وجود داشته باشد، در واقع نیازی نیست ذهن خود را در مورد تعداد پارامترهای مورد نیاز درگیر کنیم. مدل پارامتری دو مرحله دارد. مرحله اول انتخاب تابع مناسب و مرحله بعد آموزش تابع برای پیدا کردن ضرایب مناسب تابع روی داده‌های آموزش.

روش های غیر پارامتری زمانی خوب است که داده‌های زیادی داریم و هیچ دانش قبلی در مورد آن نداریم. و زمانی که نمی خواهیم بیش از حد در مورد انتخاب ویژگی‌های مناسب نگران باشیم. برای درک بهتر مدل‌های غیرپارامتری مدل k نزدیکترین همسایگی را مثال می‌زنیم. در این مدل تابع خاصی برای آموزش در نظر نمی‌گیریم و فقط سعی می‌کنیم الگوهای مشابه را در میان داده‌های آموزش پیدا کنیم.

MCC

MCC در اصل یک ضریب همبستگی بین طبقه‌بندی‌های باینری مشاهده شده و پیش بینی شده است؛ این شاخص مقداری بین -1 و + 1 را باز می‌گرداند. ضریب +1 نشان دهنده پیش بینی کامل، 0 بهتر از پیش بینی تصادفی و -1 نشان دهنده اختلاف نظر بین پیش‌بینی و مشاهده است. این ضریب در دسته‌بندی‌های دوکلاسه برای ادازه‌گیری کیفیت دسته‌بندی استفاده می‌شود.