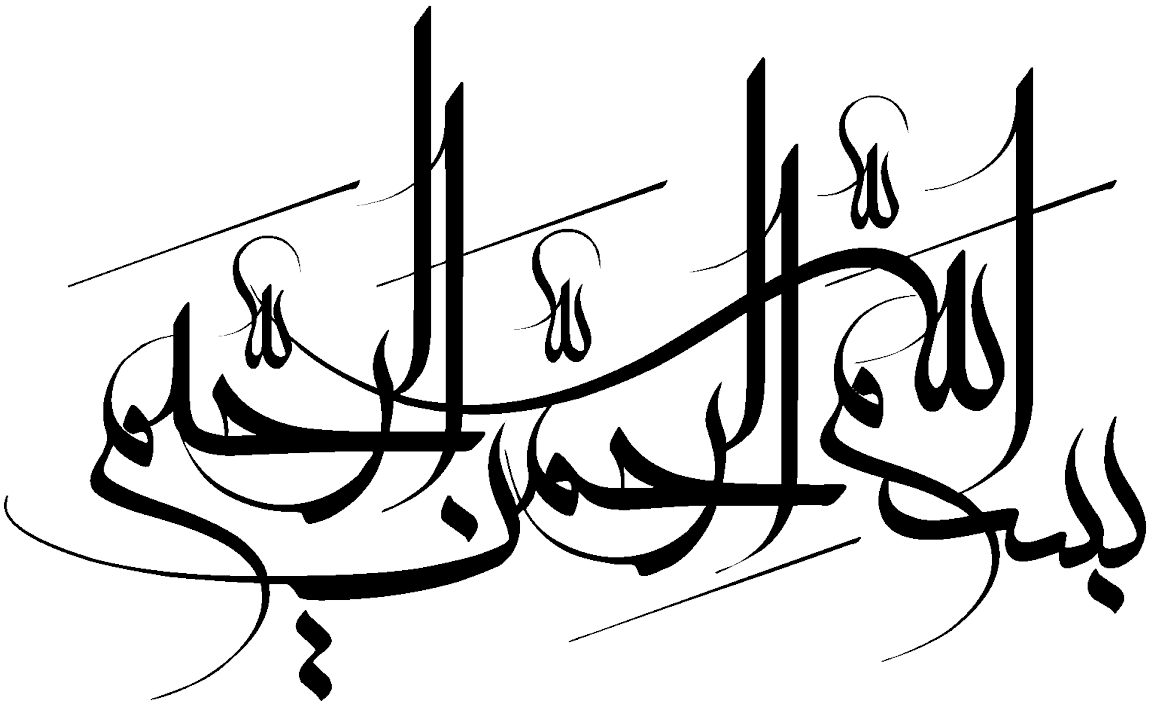


گزارش کار تمرین ۲

نام اساتید :

آقای دکتر خرد پیشه

آقای دکتر فراهانی

نام : شکیلا جابری

شماره دانشجویی :

99422046

در این تمرین data set مربوط به بیماریهای قلبی به ما داده شده است. از ما خواسته شده تا در سوال اول

Data setرا با استفاده از پکیج pandas بررسی کنیم . برای این کار باید داده ها را با توجه به data set داده شده در صورت تمرین داده ها را importکنیم که با استفاده از کتابخانه های numpy و pandas و matplotlib کتابخانه ها را فراخوانی کردیم.

در خط دوم با استفاده از فایل csv داده ها را خواندیم و با دستور head (10) ۱۰ دور اول داده ها را بدست اوردیم و دیدی نسبت به dataها را بدست اوردیم.

در خط سوم از دستور info استفاده کردیم تا بتوانیم یک شمای کلی از data setداشته باشیم و میتوانیم بفهمیم fitcher های ما چه تعداد dataهای null دارد. Dtype داده های ما چه چیزی است و اینکه چقدر از memory ما را نگهداری کرده است.

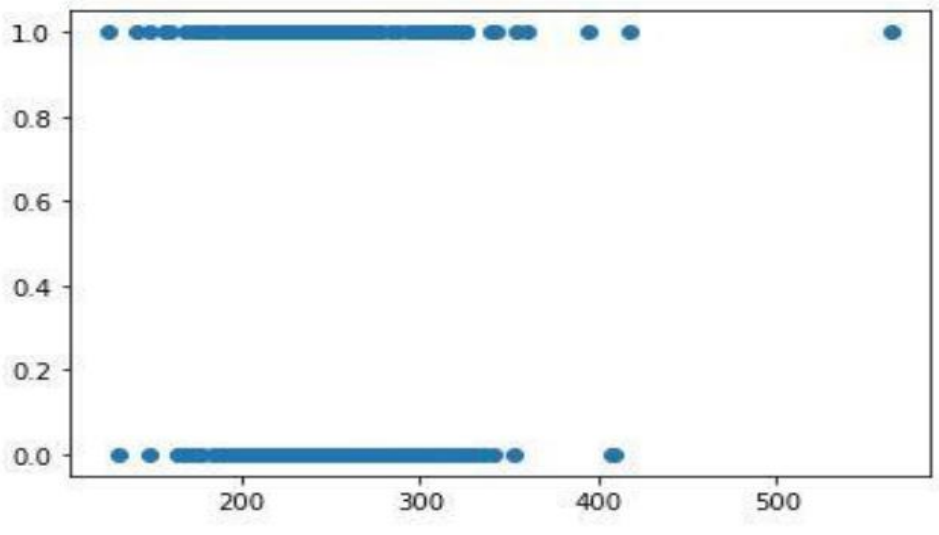
مثلا در این تمرین memory usage : 33.3 kb است.

پس تا اینجا به سوال اول پاسخ دادیم.

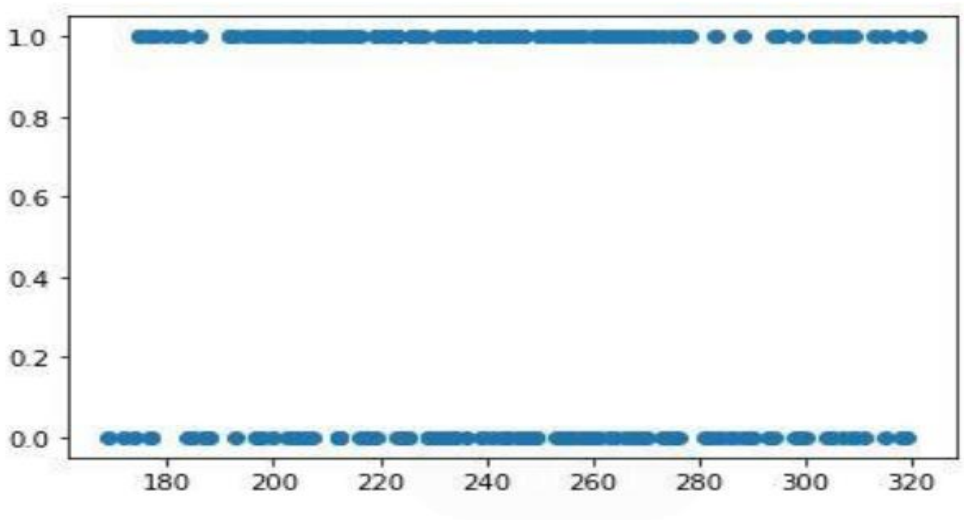
در قسمت دوم سوال از ما خواسته شده که بررسی کنیم ایا در dataset داده پرت وجود دارد یا خیر؟

با استفاده ز دستور plt scatter و با فراخ.انی datasetمتوجه میشویم که در ستون (chol)

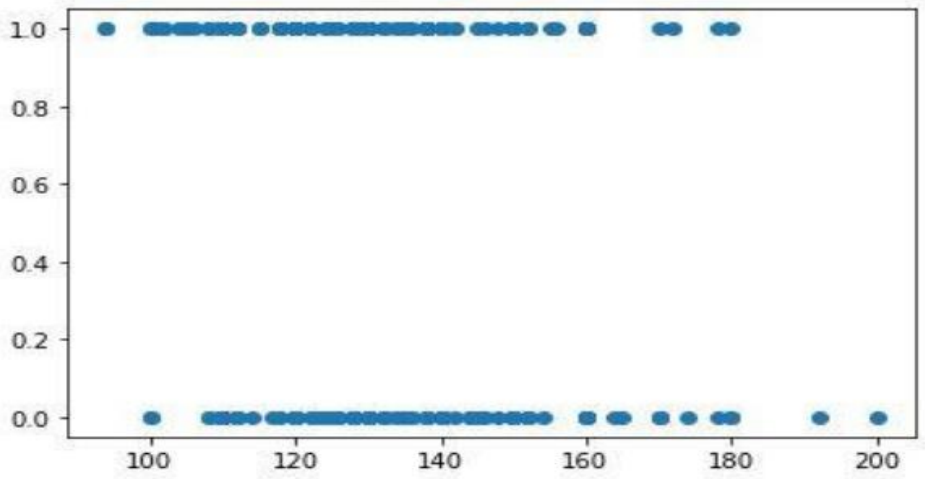
داده پرت وجود دارد و با دستور plt scatter نمودار مربوط به آن را رسم کردیم در خط چهارم.



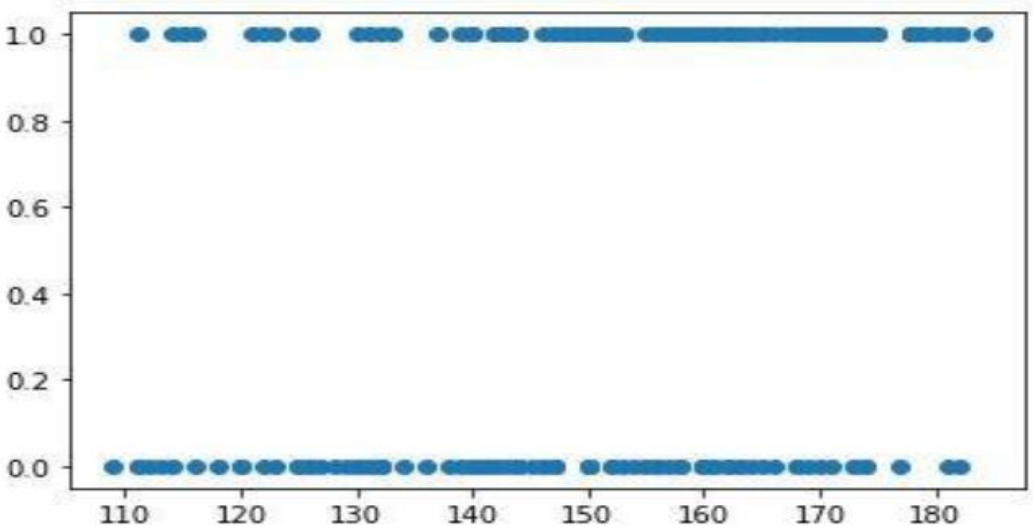
قبلا گفته بودیم که در صورت مشاهده داده های پرت باید آنها را حذف کنیم و بعد به ادامه کار بپردازیم پس داده های پرت موجود در ستون cholرا حذف میکنیم و دوباره نمودار ان را بررسی میکنیم پس در خط پنجم سعی کردیم dataهای پرت را حذف و مجدد نمودار را رسم میکنیم.نودار ستون chol بعد از حذف داده های پرت به صورت زیر خواهد بود.



در خط هفت دیتا های پرت در ستون trestbps را بررسی کردیم.



که مشاهده میکنیم دیتاهای پرت وجود دارد پس باید حذفشان کنیم.



نمودار موجود در خط ۱۲ نمودار دیتاهای ستون trestbps پس از حذف داده های پرت را نشان میدهد.

ذر سوال بعد از ما خواسته شده که بررسی کنیم که ایا تعداد نمونه ها در هر کلاس متوازن است یا خیر؟

در خط ۱۳ از دستور value_counts استفاده کردیم تا بفهمیم چه تعداد دیتاهای categorical داریم و چه تعداد دیتای عددی و string داریم.

این کار چه کمکی به ما میکنه؟ مثلا میخواهیم دیتاهای null را پرکنیم با استفاده از انها که مشاهده میکنیم که dtype ما int64 است.

در خط ۱۳ دو کلاس داریم کلاس صفر و کلاس یک.

با توجه به اطلاعات کلاسها مشاهده میکنیم که ۱۴۵ نمونه در کلاس یک و ۱۱۵ کلاس در کلاس صفر داریم.

داده ها متوازن هستند.

در اینجا از ما خواسته شده تا نمونه های موجود در dataset را با نسبت 80به 20 به دو بخش تقسیم کنیم.

در خط 14 دو متغیر x و y و از دستور axis استفاده کردیم با اینکار به هر دو بعد دسترسی پیدا کردیم.

از یک طرف به index و از طرف دیگر به colums ها که با دستور colums میتوانیم به اسم انها دسترسی داشته باشیم. مثل name team number position

در خط 15 با استفاده از پکیجی که در صورت سوال خواسته شده یعنی sklearn . data set را به دو بخش داده اموزشی وقسمت تقسیم میکنیم.

پس این کتابخانه را import میکنیم.

در سوال سوم از ما خواسته شده تا قضیه بیز را بیان کنیم قابل بیان است که روشی است برای دسته بندی پدیده ها به پایته احتمال وقوع یا عدم وقوع یک پدیده.

اگر برای فضای نمونه ای مورد نظر بتوانیم چنان افرازی انتخاب کنیم با داشتن اینکه پیش امد های افراز شده رخ داده بخش هایی از عدم قطعیت کم میشود.

از طریق این قضیه میتوان احتمال یک پیشامد را با مشروط کردن نسبت به وقوع یک پیشامد دیگر محاسبه کرد.

در ادامه سوال خواسته شده که دسته بندی های Gaussian وnave bay es multinomial naive naïve beyes, berroulli beyes را با یکدیگر مقایسه کنیم و کاربدرشان در ذکر کنیم.

به خط ۱۶ رجوع میکنیم .

در این خط برای بررسی و مقایسه متد کتابخانه ای multi variable normal را فراخاونی کردیم.

تحلیل واریانس چند متغیره یا manova :

این تحلیل یکی از پیچیده ترین ازمونهای اماری است. تحلیل واریانس یکطرفه anova برای ازمون مقایسه میانگین یک متغر کمی در بین بیش از دو گروه مستقل استفاده میشود. در واقع manovaتعمیم یافته ازمون T است. ور دارای همان پیش فرض هاست و تنها تفاوت این است که میانگین متغیرهای کمی در بیش از دو گروه مستقل با هم مقایسه میشوند. وظیفه اصلی این ازمون این است که بطور همزمان بیا کند که مولفه ها بصورت یکجا ایا در بین گروه های مسقل متغیرهای کمی تفاوت میانگین دارد یا خیر؟

به جای استفاده از این ازمون چندبار از ازمون anova استفاده نمیکنیم؟ ازمون anova به بررسی تفاوت میانگین یک متغیر کمی در گروه های یک متغیر کمی میپردازد.

پس میانگین واریانس را مقایسه کردیم در Gaussian .

با استفاده از دستور len میتوانیم بفهمیم چه تعداد نمونه وجود دارد.میت.انیم از دستور shapeهم استفاده کنیم که در ادامه استفاده کردیم تا تعداد ستونها را مشخص کنیم و به ما اطلاع میدهد که باید یکسری کار اماری انجام دهیم .

در کد ۱۶ در خط def fit (self,x,y) و def predict (self,x,gussian) یک کلاس تعریف شده که ۲ کلاس دارد.

در تابع اول ۳ پارامتر تعریف شده است.

در خط selfing gussian = dict() در خط اول دیکشنری جدید در گوس پارامتری که در پایین پر میکنیم قرار میگیرد.

در خط self.perios=dict در خط دوم دیکشنری جدید در بخش پیشین پارامتر که در پایین پر میکنیم قرار میگیرد.

در خط lables = set(y) یک متغیری تعریف میکنیم و در داخلش مجموعه ای از Y قرار میدهیم.

در خط for c in lables یک حلقه ایجاد میکنیم که به تعداد متغیر بالا کار میکند.

در خط current x=X(Y=c) در این خط یک متغیر تعریف شده است .در داخل آن یک ارایه قرار دارد و اندیس این ارایه زمانی درست است که مقدار پارامتر با مقدار بعدی برابر باشد.

و در خط self.gussian در این بخش به تعداد اندیس داده ها را داخل پارامتر قرار میدهیم.

و در خط return یک تابع گوس رو برمیگردانیم.

اگر داده ها از نوع پیوسته باشد از مدل احتمالی با توزیع گوسی یا نرمال برای متغیرهای مربوط به شواهد میتوان استفاده کرد.

بیز چند جمله ای به عنوان یک دسته بندی متنی بسیار به کار می اید.

در این حالت برحسب مدل احتمالی یا توزیع چندجمله ای برداری از n ویژگی برای یک مشاهده به صورت X=(x1,...,xn) با احتمالات (p1,...,pn) در نظر میگیریم پس بردار x نشانگر تعداد مشاهداتی است که ویژگی خاصی دارند.

نایوبیز برنولی بیشترین کاربرد را در دسته بندی متنهای کوتاه دارد.

در خط shape b وd=x وN دو متغیر تعریف شده که در داخل هر متغیر یک مقدار از نوع عدد قرار میگرد.

در خط def predict (self,x) در این خط تابع ها ۳ پارامتر تعریف شده است.

در خط k=len(self) دراین خط یک متغیر تعریف میکنیم و در داخل آن طول گوسی را قرار میدهیم.

در سوال ۴ از ما خواسته شده تا با در نظر گرفتن فیچرها یک دسته بند Gaussian naive را پیادهسازی کنیم.

در خط ۱۷ فیچرهای trestbps-thalach-chol را در یک دسته بند قرار داده ایم.

در خط ۱۹ با اسفاده از پکیج sk learn توانستیم خروچی را بدست اوریم. مدل نایو بیز گایوسی بدون استفاده از کتابخانه ها ساختیم.