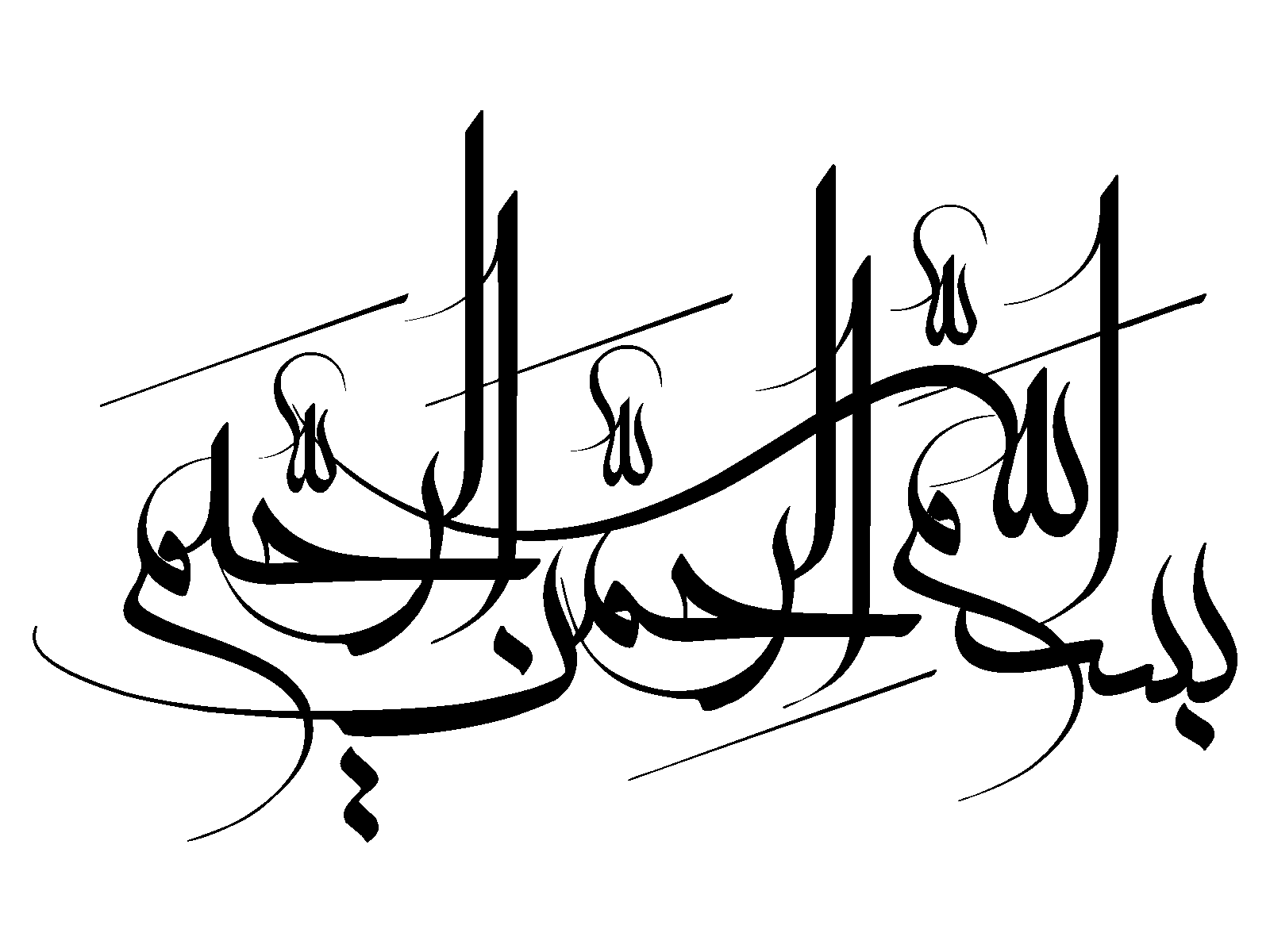
ا

original

[Draw your reader in with an engaging abstract. It is typically a short summary of the document. When you’re ready to add your content, just click here and start typing.]

[Document title]

[Document subtitle]



گزارش کار تمرین 2

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 import کنیم 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 دستورhead (10) 10 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 set داشته باشیم و میتوانیم بفهمیم feature های ما چه تعداد dataهای null دارد. Dtype داده های ما چه چیزی است و اینکه چقدر از memory ما را نگهداری کرده است.

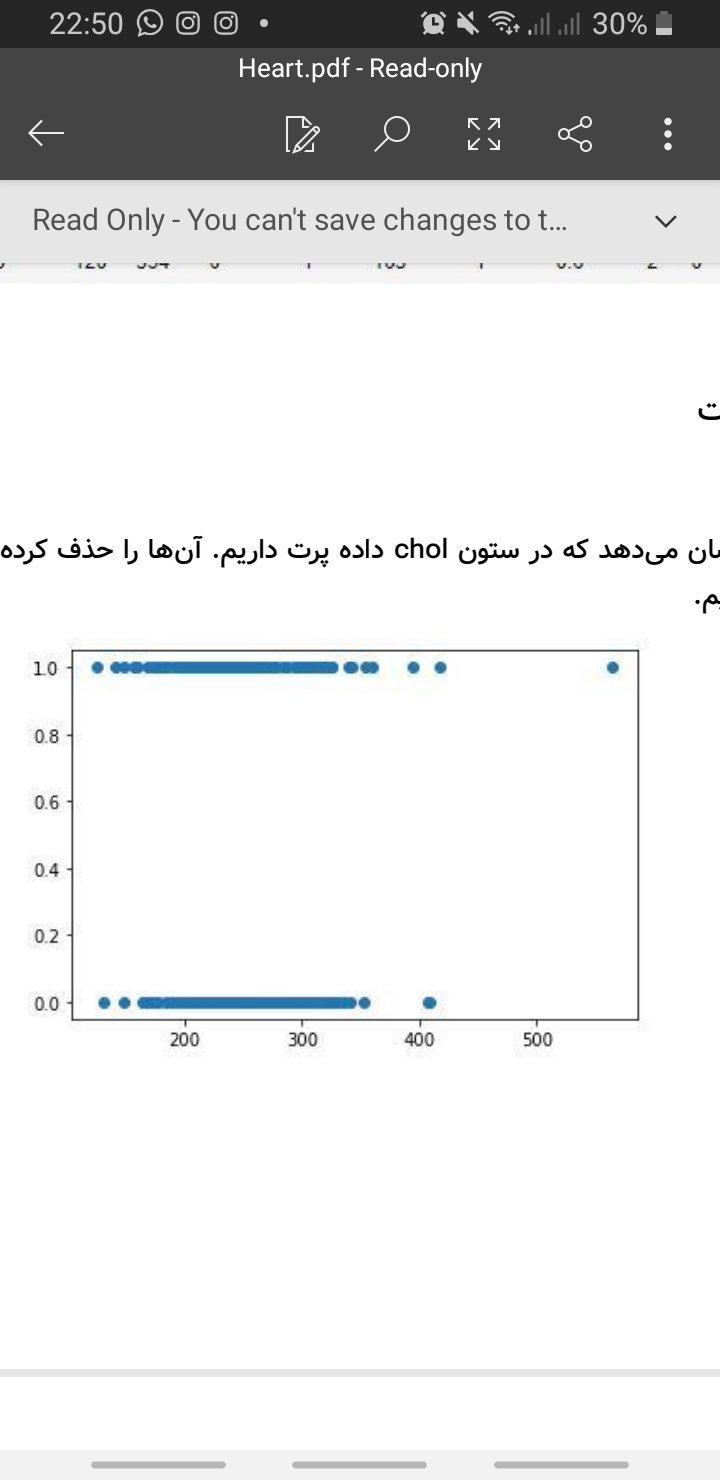
مثلا در این تمرین memory usage : 33.3 kb است.

پس تا اینجا به سوال اول پاسخ دادیم.

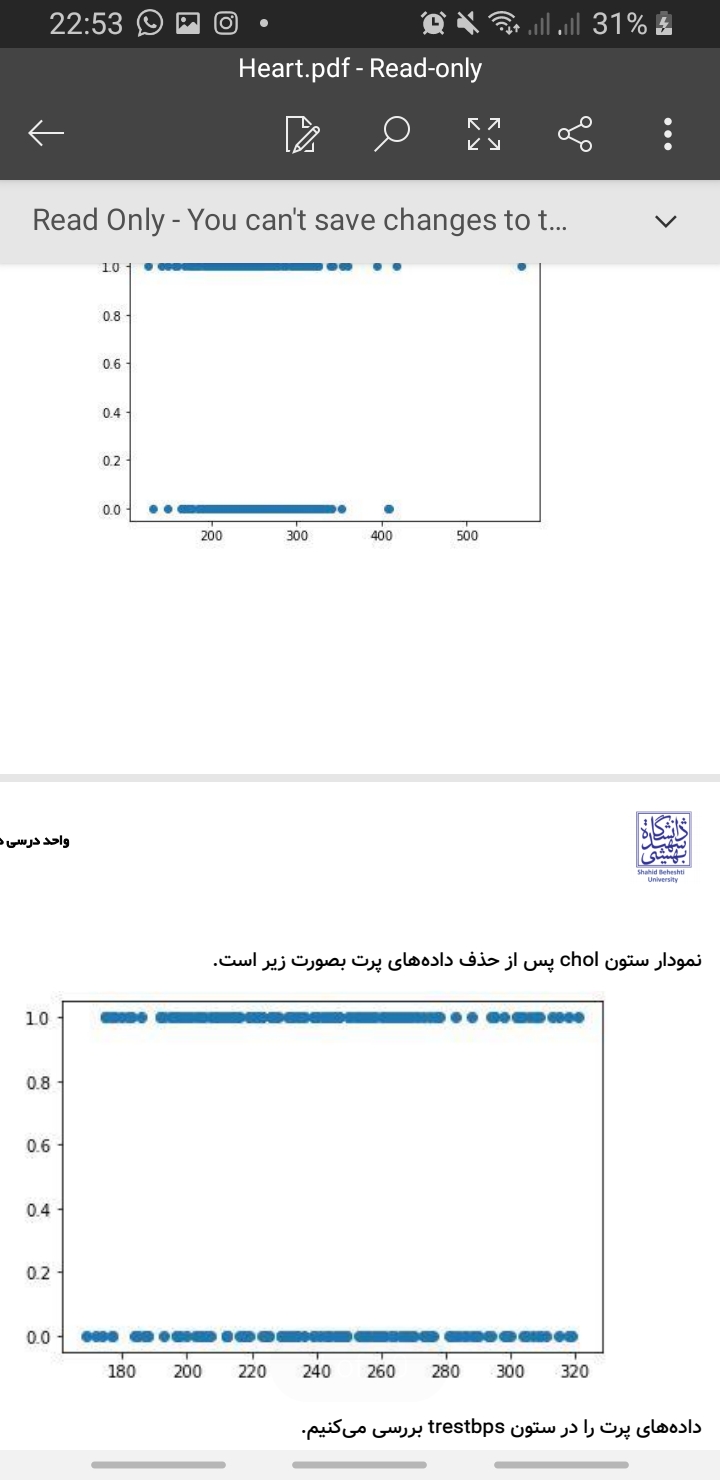
در قسمت دوم سوال از ما خواسته شده که بررسی کنیم ایا در dataset داده پرت وجود دارد یا خیر؟

با استفاده ز دستور scatter plt و با فراخوانی dataset متوجه میشویم که در ستون (chol)

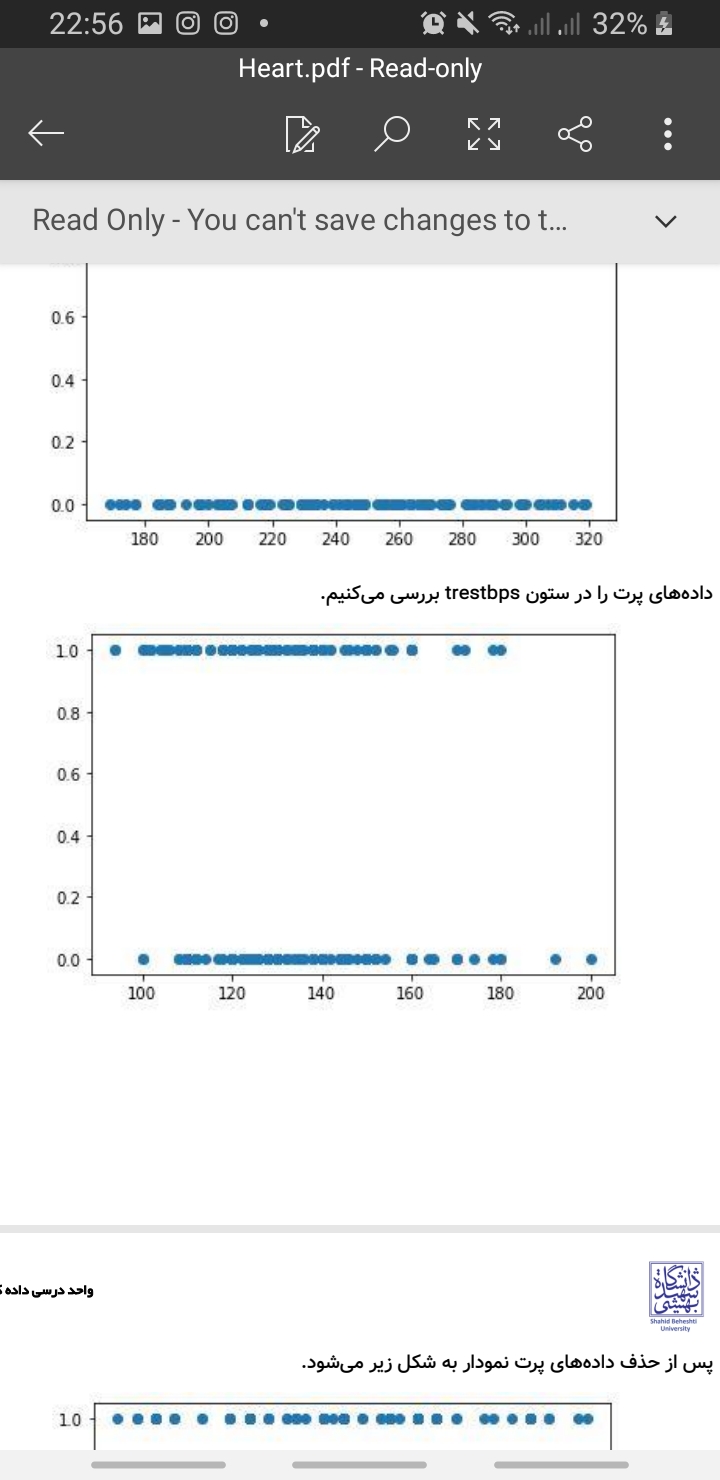
داده پرت وجود دارد و با دستور plt scatter نمودار مربوط به آن را رسم کردیم در خط چهارم.



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 chol را 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 12 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 13 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 13 دو کلاس داریم کلاس صفر و کلاس یک.

با توجه به اطلاعات کلاسها مشاهده میکنیم که 145 نمونه در کلاس یک و 115 کلاس در کلاس صفر داریم.

داده ها متوازن هستند.

در اینجا از ما خواسته شده تا نمونه های موجود در dataset را با نسبت 80 به 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 team number position name

در خط 15 با استفاده از پکیجی که در صورت سوال خواسته شده یعنی sklearn . data set را به دو بخش داده اموزشی وتست 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 پایتون 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 و nave bay es multinomial naive berroulli beyes naïve beyes, را با یکدیگر مقایسه کنیم و کاربردشان در ذکر کنیم.

به خط 16 رجوع میکنیم .

در این خط برای بررسی و مقایسه متد کتابخانه ای multi variable normal را فراخوانی 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 manova تعمیم یافته ازمون T است. و 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 چرا 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.میتوانیم از دستور shape هم 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 16 در خط def fit (self,x,y) و def predict (self,x,gussian) یک کلاس تعریف شده که 2 کلاس دارد.

در تابع اول 3 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 و d=x و N 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 3 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 4 از ما خواسته شده تا با در نظر گرفتن فیچرها یک دسته بند Gaussian naive را پیادهسازی کنیم.

در خط 17 فیچرهای trestbps-thalach-chol را در یک دسته بند قرار داده ایم.

در خط 19 با اسفاده از پکیج sk learn توانستیم خروچی را بدست اوریم. مدل نایو بیز گایوسی بدون استفاده از کتابخانه ها ساختیم.

خط بیستم: from skleran .naive bayes import Gaussian nb در این خط کتابخانه را import کرده ایم

Gnb1-clf-gaussian nb: مدل naïve bayes را به یک متغیر میریزیم

Gnb1-clf .fit (x-train [chol,thalach,…),y-train

در این قسمت با تابع fit دو تا ارگومان قرار گرفته است در ارگومان اول که x-train هست چندتا از نام فیلد از data set نوشته شده و در ارگومان دوم هم فقط متغیر قرار دادیم

خط بیست و یکم: prediction = gnb1-clf.predict…..

در این قسمت با متغیر gnb1-clf و با تابع predict یک ارگومان دریافت کرده و چند فیلد هم در داخل x-test قرار داده و تمام این عملیات را در داخل یک متغیر به نام prediction قرار داده ایم

خط بیست و دوم: در این قسمت عملیات چاپ را انجام داده ایم با دستور print

خط بیست و سوم: from skiearn .svm import svc

در این قسمت هم کتابخانه را import کردیم

Svm-clf=svc

در این قسمت مدل svc را داخل یک متغیر قرار میدهیم

توضیحی در موردsvm:

Support vector machins: در ابتدا svm را برای مساِیل classification دو کلاسه استفاده میکنیم و بعد به چند کلاسه تعمیم میدهیم

در svm هدف این است که یک خط پیدا کنیم که بتواند dataها را از هم جدا کند یعنی میخواهیم دیتا را separate کنیم

به وسیله یک لاین که این خط حالت خاصی از مفهوم hyper plain است

Feature space چیست: فرض میکنیم دو تا فیچر داریم x1 و x2

مثلا x1در کلاس بنفش قرار دارد و x2در کلاس ابی

ما میخواهیم کاری کنیم که 2 منطقه ابی و بنفش به وسیله ی یک line از یکدیگر جدا شوند

گاهی اوقات ما میتوانیم به وسیله یک خط دیتا را جدا کنیم اما گاهی نمیتوانیم .

همیشه یک تعدادی از data بیرون خواهد افتاد پس باید راه حل دیگری به کار ببریم.

Dataهایی که نمیتوان از یکدیگر جدا کرد خطی جدایی پذیر نیستند

کاری که در svmمیخواهیم انجام دهیم این است که dataها را به کمک یک hyper plain جدا کنیم پس یک فضایی داریم که به وسیله یک hyper plain میخواهیم سعی کنیم دیتا ها کلاس بندی کنیم.

1-حالا در بعضی داده ها dataها کاملا خطی جدایی پذیرند

2-در برخی دیگر میتوانند جدا شوند اما خطاهایی داریم

3-در برخی دیگر هم اصلا خطی جدایی پذیر نیستند

در نوع اول گفتیم با استفاده از یک هایپر پلین این کار را انجام میدهیم و عملا فضای فیچرهایمان را بزرگتر میکنیم

در واقع در svmخطی را در نظر میگیریم که بیشترین مارجین یا حاشیه را داشته باشد.

خط بیست و چهارم: در این خط فقط نوع کتابخانه تغییر کرده است

خط بیست و پنجم:بازهم با استفادهد از دستور پرینت عملیات چاپ کردن را انجام میدهیم

حط بیست و ششم : svm-clf=svc (kernel =rbf)

اگر یک فراخوان داده شود از آن برای پیش محاسبه اصلی ماتریس از ماتریس داده استفاده میشود.

آن ماتریس باید یک آراتیه از شکل باشد .

مجددا عملیات چاپ را انجام میدهیم

در این بخش از کتابخانه svc استفاده کرده ایم.

خط بیست و نهم:svm-clf=svc (kernel =sigmoid)

در این خط svc نوع هسته را برای استفاده در الگوریتم مشخص میکند .

Sigmoidنوع kernelرا مشخص میکند.

اگر یک کالایی عرضه شود از آن برای پیش محاسبه ماتریس هسته از ماتریس داده استفاده میشود

آن ماتریس باید یک ارایه از شکل باشد .

خط سی وسی و یکم پیش تر توضیح داده شده است و باز هم با دستور printاطلا:عات را چاپ کرده ایم .

خط سی و دوم: svm-clf =svc (kernel=(poly))

در اینجا نوع متغیر را poly قرار داده ایم و باز هم عملیات چاپ را انجام داده ایم